Clamart : EDF – Collection de notes internes de la Direction des Etudes et Recherches, Rapport n° 97NO00011
© EDF 1997 – ISSN 1161-0603

Etat de l'art des idées implémentées dans les moteurs de recherche par index sur WWW

State of the art of the ideas implemented in search engines on the WWW

Bénédicte PINCEMIN, Xavier LEMESLE

Résumé

Une dizaine d'outils de recherche par index sur le World Wide Web sont étudiés, non pour en faire un comparatif, mais pour tirer un état de l'art des modélisations utilisées. Les modes de représentation des documents et des requêtes, de calcul de la pertinence, et de visualisation des résultats, sont disséqués pour répertorier tous les paramètres considérés et toutes les propriétés des textes exploitées.

Cet inventaire structuré peut alors servir de base à la conception d'un nouvel outil de recherche en texte intégral, tirant parti de toutes les « bonnes idées » de systèmes opérationnels et reconnus. Il fournit également une grille d'analyse approfondie de tout moteur de recherche documentaire actuel, voire un référentiel pour un comparatif.

La Note s'adresse donc aux concepteurs de systèmes de recherche documentaire, ainsi qu'aux utilisateurs qui veulent approfondir leurs connaissances du fonctionnement des outils pour mieux les maîtriser et en tirer le meilleur parti. En complément, une sélection de références disponibles sur WWW sont présentées.

Cette étude se fonde sur des observations réalisées entre août et octobre 1996, dans le cadre du projet DECID (Diffusion Electronique Ciblée d'Informations et de Documents).

Executive summary

Ten search engines indexing the World Wide Web (with keywords or concepts) were studied, not in order to make any comparison between them, but to obtain an idea of the state of the art of the models being used. The way in which documents and requests are represented, the calculation of the relevance and the results visualisation, were all dissected in order to list all the parameters under consideration and all the features of the texts that were examined.

This structured inventory may therefore be used as a basis for designing a new full-text retrieval system, using all the "« good ideas » that were found in the above-mentioned tools, all of which are operational and popular. The inventory also provides a sophisticated frame for analysing all current information retrieval tools, and can be used as a reference for comparison.

This technical report is therefore intended for information retrieval system designers and users who wish to improve their knowledge about how such tools work, in order to get more out of them. A selection of the references available on the WWW is also supplied.

This study is based on observations made between August and October 1996, as part of the DECID project (Diffusion Electronique Ciblée d'Informations et de Documents = targeted electronic distribution of informations and documents). The DECID team is building a document routing system for selective dissemination of information throughout EDF Research Centre).

Synthèse

Les principaux services de recherche sur WWW (AltaVista, Excite, Harvest HomePage Broker, HotBot (ex-Inktomi), Infoseek Guide, Lycos, Magellan, Open Text Index, WebCrawler, WWW Worm) sont analysés, à partir de la façon dont ils se présentent à l'écran, de leur description (notamment dans l'aide en ligne), et de leur comportement. Ceci a permis de faire une typologie des paramètres jugés utiles (par les concepteurs et les utilisateurs), avec la manière de les utiliser efficacement. Quatre phases sont à prendre en compte, chacune étant déterminante pour la qualité des résultats :

- *la constitution de la base*, avec la description interne des documents et notamment de leurs parties textuelles ; il faut souligner que l'appréciation d'un service dépend considérablement de la qualité de ses données et de sa couverture.
- *le mode de soumission et la mise en forme d'une requête*. Il y a deux choix de conception : quelles formes peuvent prendre les termes élémentaires de la requête, et comment ils sont articulés. En ce qui concerne la forme des termes, il faut considérer d'une part la gestion des variations linguistiques (lemmatisation et réduction dérivationnelle en morphologie et syntaxe, réduction conceptuelle en sémantique), et d'autre part l'apport de pouvoir faire des recherches sur des syntagmes (mots composés, expressions). Quant à l'articulation ensuite des termes composant la requête, l'utilisation d'une syntaxe formelle (par exemple booléenne) peut être vue comme contraignante (par les tenants des interfaces en langue naturelle) ou au contraire comme permettant une interrogation plus précise et plus puissante (combinaison d'opérateurs, pondérations). Un juste équilibre est à trouver, pour respecter les habitudes d'interrogation sans pour autant niveler par le bas (en s'interdisant des options plus complexes mais efficaces).
- *l'évaluation de la pertinence*. La pertinence est de moins en moins comprise comme une grandeur binaire (« c'est (tout-à-fait) pertinent, ou ça ne l'est pas (du tout) ») ; on lui associe plutôt une valeur sur une échelle continue, traduisant un jugement gradué (« c'est plus ou moins pertinent »). Le calcul de la pertinence combine en général des données sur la fréquence, la position, et le caractère discriminant des termes de la requête, et sur la proximité des occurrences des uns et des autres.
- *la présentation des résultats* et l'ergonomie du système. Le volume de documents potentiellement pertinents est souvent énorme. Pour que le résultat soit exploitable, il faut : que l'utilisateur puisse ajuster le volume de références à dépouiller (seuillage, tri) ; qu'il puisse aisément se faire une idée du contenu de chaque document proposé ; qu'il comprenne ce qui a motivé la sélection d'un document par le système (notamment pour mieux ajuster sa requête) ; qu'il ait des éléments pour estimer la fiabilité et la représentativité des réponses obtenues ; qu'il puisse garder trace de ses recherches antérieures. La qualité de la documentation en ligne du système compte aussi beaucoup pour la bonne interprétation des résultats et la (re)formulation de la recherche. Elle est malheureusement très inégale, la dissimulation plus ou moins ouverte des algorithmes de traitement étant monnaie courante.

Trois fonctionalités avancées retiennent particulièrement notre attention : l'utilisation d'un document comme requête (Excite, Infoseek Guide, WebCrawler), la recherche par concept (Excite), et l'utilisation de pondérations (Excite, HotBot). Dans tous les cas, les principes sont intéressants, mais les implémentations actuelles ne donnent pas pleinement satisfaction. D'abord, il y aurait à gagner à bien distinguer (dans le traitement) les requêtes sous forme de mots-clés de celles sous forme de textes. Ensuite, nous dénonçons l'amalgame que nous observons entre le QBE (Query-By-Example) et le relevance-feedback (retours utilisateurs). De plus, la recherche « par concepts » doit être démystifiée, pour engager une réflexion sur la définition et la construction des classes thématiques sous-jacentes et sur leur mode d'utilisation. Enfin, quand

l’utilisateur peut indiquer lui-même l’influence des termes par l’intermédiaire de pondérations, les systèmes de pondération numérique (chiffrée) sont peu puissants et portent à confusion, et les systèmes plus qualitatifs restent à développer.

En définitive, les services de recherche par index sur WWW ne mettent pas en œuvre de techniques originales ou innovatrices, par rapport à ce qui était connu en matière de recherche documentaire informatisée. Cet état de l’art permet néanmoins de faire un point complet des idées efficaces, implémentables à grande échelle, et opérationnelles.

Sommaire

1.MOTIVATION ET ORIENTATION DE L'ENQUÊTE **5**

1.1.CONTEXTE : LES MOTEURS DE RECHERCHE SUR WWW, DES OUTILS APPRÉCIÉS 5
1.2.BUT : UN ÉTAT DE L'ART DES IDÉES SOUS-JACENTES AUX OUTILS, PLUTÔT QUE DES OUTILS EUX-MÊMES 6
1.3.MÉTHODOLOGIE EMPLOYÉE : UNE ENQUÊTE MINUTIEUSE À PARTIR DES DOCUMENTATIONS DISPONIBLES SUR WWW 7

2.RÉSULTAT DE L'ENQUÊTE : RÉCAPITULATIF SYSTÉMATIQUE DES PARAMÈTRES **9**

3.POINTS D'APPROFONDISSEMENT **26**

3.1.L'UTILISATION DU DOCUMENT COMME REQUÊTE : QBE, RELEVANCE-FEEDBACK 26
3.2.LA RECHERCHE PAR CONCEPTS 28
3.3.LES MODES DE PONDÉRATION : PONDÉRATIONS QUANTITATIVE ET QUALITATIVE 29

4.CONCLUSION **32**

5.BIBLIOGRAPHIE-WEBLIOGRAPHIE THÉMATIQUE COMMENTÉE **33**

5.1.COMPARATIFS : ÉVALUATION DES SYSTÈMES 33
5.2.CADRE D'ANALYSE ET CRITÈRES GÉNÉRAUX D'ÉVALUATION 34
5.3.SUR LES TECHNIQUES DE CALCUL IMPLÉMENTÉES 34

1. Motivation et orientation de l'enquête

1.1. Contexte : Les moteurs de recherche sur WWW, des outils appréciés

Le foisonnement de documents accessibles sur WWW a suscité la conception et la mise à disposition[1] d'outils de recherche. Pour ne citer que quelques-uns parmi les plus connus aujourd'hui : ALIWEB, ALTAVISTA, EXCITE, GALAXY, HARVEST HOMEPAGE BROKER, HOTBOT (ex-INKTOMI), INFOSEEK GUIDE, LYCOS, MAGELLAN, OPEN TEXT INDEX, WEBCRAWLER, WWW WORM, YAHOO...

Nous nous intéressons ici à ces outils de recherche pour plusieurs raisons :

- ce sont des outils effectifs, utilisés, reconnus : donc pour nous à la fois des exemples d'applications réelles (non des prototypes ni des spéculations théoriques), et des points de repère par rapport auxquels situer notre système DECID[2] ;
- ils sont conçus pour traiter de grands (gigantesques...) corpus textuels, sur les sujets les plus variés, avec sans cesse l'arrivée de nouveaux documents et l'impératif d'avoir une information aussi à jour que possible : ceci correspond à des approches robustes et fortement automatisées, telles celle que nous avons à mettre en place.

Pour être précis, deux grandes classes d'outils de recherche doivent être distinguées :

- les outils « type index » basés sur une indexation des documents : l'utilisateur soumet au système l'objet de sa recherche sous forme d'une requête. Le système propose alors des documents susceptibles d'y répondre sur la base de mots-clés les caractérisant.
- les outils « type répertoire » donnant accès aux documents par l'intermédiaire d'un catalogue thématique : chaque document est préalablement classé dans une (ou plusieurs) rubrique(s) ; la recherche d'une information se fait en parcourant l'arborescence du plan de classement, en partant des rubriques générales et en cernant peu à peu le sujet concerné.

Nous faisons le choix de nous intéresser uniquement aux outils basés sur l'indexation.

En effet, pour diverses raisons que nous ne développerons pas ici, l'autre formule, celle d'un catalogue thématique, ne convient pas dans l'optique du système DECID. Pour mention, voici néanmoins un aperçu des arguments qui peuvent opposer index et répertoire dans certaines applications, avec les points en défaveur de la forme répertoire :

- la conception normative de l'espace de recherche : la structure est fixée, et impose un « découpage » particulier de la réalité à décrire. Elle est tôt ou tard confrontée à des

[1] Les outils généralistes sont d'utilisation gratuite ; ils vivent des publicités insérées dans leurs pages. Ils font aussi parfois office de démonstration d'une technologie par ailleurs pouvant être acquise, moyennant finances, pour gérer des fonds documentaires privés.
En revanche, des services professionnels plus spécialisés -commerce, finances, technique- monayent l'accès à leur données.
INFOSEEK GUIDE par exemple décline ces deux types d'offre.
Dans l'ensemble, les services de recherche tendent en fait à perdre leur caractère désintéressé : les outils initialement proposés par des équipes de recherche universitaires ont été rachetés, ou sont cédés via des licences. Pour que les intérêts commerciaux ne gouvernent pas l'ensemble des services du WWW, des initiatives publiques se font jour, cf. par exemple le projet européen DESIRE [Koch].

[2] DECID est un acronyme pour « Diffusion Electronique Ciblée d'Informations et de Documents ». Tout agent responsable d'ARD/AID à la DER est caractérisé par un profil défini automatiquement à partir de textes descriptifs de son activité (textes d'ordonnancement). Un calcul de similarité entre textes permet alors de retrouver les agents a priori les plus concernés par un document donné. La même technique est appliquée pour constituer le « Qui-Fait-Quoi ? » annuel. Les trois piliers fondamentaux qui caractérisent l'approche DECID sont : (i) l'*automatisation* de la construction des profils et des calculs de proximité ; (ii) la caractérisation des profils, aussi bien que des documents, à partir de *texte intégral* ; (iii) la vision des profils formant une *base*, et non pas existant indépendamment les uns des autres.

documents atypiques, qui soit sont exclus, soit sont mis à part dans une classe « divers » difficilement exploitable.

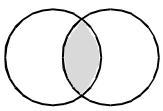
- l'objectivisme qu'il y a à analyser *a priori* et une fois pour toutes les documents : on ne tient pas compte du point de vue porté à travers une recherche, qui modèle, réorganise, réévalue l'ensemble des documents.
- le caractère figé du plan de classement hiérarchique, qui laisse prévoir sa caducité.
- le partitionnement des documents en une hiérarchie de classes d'équivalences, avec la difficulté et l'arbitraire qu'il y a pour déterminer le découpage en niveaux et (conjointement) le rapport d'équivalence qui lie les documents.
- l'inadéquation à une trop grande masse d'information, qui conduit d'ailleurs la plupart des systèmes existants à sélectionner les documents qu'ils intègrent. En effet, des limitations d'ordre cognitif font qu'on ne peut raisonnablement démultiplier ni les niveaux de classement, ni les options à chaque niveau, ni le nombre de documents dans chaque classe : tout parcours a ses limites.

1.2. But : Un état de l'art des idées sous-jacentes aux outils, plutôt que des outils eux-mêmes

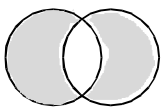
Soyons donc bien clairs sur les objectifs de cette étude.

Il ne s'agit pas pour nous de faire un banc d'essais des outils existants, de façon à discerner « le meilleur » d'entre eux, ou, de façon plus réaliste (car il n'y a pas d'outil « idéal » et universel !), à dégager des caractéristiques suffisantes pour guider le lecteur vers tel ou tel outil selon le type de recherche à effectuer et le type de résultats attendus. (Les outils étant complémentaires, la solution la plus satisfaisante consistera d'ailleurs parfois à combiner les résultats de plusieurs d'entre eux.) Celui qui a besoin de ce genre de comparatifs a tout intérêt à consulter les documents WWW sur le sujet : il en existe déjà de nombreux, avec toutes les variantes de points de vue et de niveau de détail. Nous recommandons en bibliographie ceux que nous avons le plus appréciés pour leur sérieux, leur caractère systématique, et parfois leur originalité (sur le plan des renseignements apportés et des remarques). Consulter WWW convient tout-à-fait pour trouver les comparatifs les plus à jour ; on en trouve aussi dans les revues spécialisées (*PC Expert, Internet World*...), mais certains de ces articles sont également disponibles sur WWW. Il faut être conscient que de telles études sont extrêmement vite caduques car les moteurs ne cessent d'évoluer. En effet, une bonne idée qui serait spécifique à l'un des moteurs (par exemple pour le calcul des documents pertinents ou pour l'interface) a toutes les chances d'être reprise et adoptée par les autres, mettant ainsi en défaut le comparatif qui se voulait discriminant.

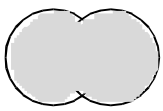
Dit de façon imagée, ces bancs d'essais s'intéressent à l'*intersection* (les fonctionalités communes) et surtout à la *différence symétrique* (les fonctionalités originales) entre les ensembles de caractéristiques associées à chaque outil. Notre approche, elle, serait de cerner l'*union* de toutes ces caractéristiques :



intersection (∩)



différence symétrique (Δ)



union (∪)

Point de vue des comparatifs

Notre point de vue

Notre but est d'abord de recenser, aussi exhaustivement que possible, toutes les méthodes et les fonctionalités implémentées et proposées par les moteurs de recherche actuels. Puis nous voulons les organiser pour en présenter une synthèse qui les regroupe logiquement et les résume.

C'est en quelque sorte un triple état de l'art :

1. des techniques de caractérisation des documents,
2. des paramètres combinés dans le calcul de leur pertinence par rapport à une requête,
3. de la manière de répondre aux attentes et aux besoins de l'utilisateur par une ergonomie appropriée.

L'enjeu n'est pas d'analyser tous les outils de recherche existants (il y en aurait des dizaines et des dizaines), ceux qui sont intéressants pour nous ici sont ceux qui sont effectivement utilisés et donc aussi les plus connus.

Nous avons donc tout particulièrement examiné : ALTAVISTA, EXCITE, HARVEST HOMEPAGE BROKER, HOTBOT (ex-INKTOMI), INFOSEEK GUIDE, LYCOS[3], MAGELLAN, OPEN TEXT INDEX, WEBCRAWLER, WWW WORM[4].

1.3. Méthodologie employée : une enquête minutieuse à partir des documentations disponibles sur WWW

Nous avons recherché sur WWW tous les renseignements possibles sur les caractéristiques des moteurs de recherche et les techniques employées.

Notre première source d'information sont les documentation-utilisateur (aides en ligne, sous les rubriques « Help ») pour chaque outil.

Bien sûr, pour garder avantage sur ses voisins, nul ne veut complètement dévoiler ses algorithmes ! L'information technique est donc généralement incomplète, morcelée, formulée de façon évasive et qualitative[5]. En nous appuyant sur notre connaissance des techniques classiques de recherche documentaire par des moyens de statistique et de linguistique computationnelle, nous avons en partie décodé cette information lacunaire pour en tirer les principes et les paramètres sur la base desquels sont faits les calculs.

Quelques documents complémentaires, notamment des comparatifs de moteurs de recherche, nous ont aidé dans cette tâche de description et d'interprétation : pour la description, grâce à l'inventaire systématique des caractéristiques qu'ils proposent (leur grille d'analyse des moteurs) ; pour l'interprétation, via les observations qu'ils ont notées et les renseignements qu'ils ont recueillis dans des articles informés ou auprès de concepteurs.

Tout repose donc sur une collection d'indices et une bonne connaissance du domaine applicatif.

Ceci serait-il une extrapolation ambitieuse ou audacieuse ? Peut-être pas tant que cela : il nous importe moins de retrouver les formules de calcul exactes (qui, de toutes façons, restent un

[3] La présentation de LYCOS s'est considérablement dégradée au cours de la rédaction de cette étude. Au début, nous avons pu trouver quelques précieux renseignements (cf. **IV.B.1.b.** et note à propos du seuillage), en particulier dans les FAQ (cf. **IV.B.2.b.**). Maintenant, toute la documentation s'est réduite à une peau de chagrin, concrètement deux pages fanfaronnes sans réelle valeur informative. Car où est donc la prétendue aide contextuelle ? :
« Do not be alarmed : on the LYCOS site, you will not find a « FAQ ». What you will find in place of the highly impersonal, dissatisfying « frequently asked questions » list, is a feature we understately call « Help ». At any time, click on the « Help » button (on the left menu of any LYCOS screen) and you'll find tips on getting what you want immediatly -based on where you are in the site. We like to think of it as less FAQ, more YAQ (« your answered questions »). »

[4] Il est vrai que WWW WORM fait figure d'ancêtre des moteurs de recherche sur WWW, et ne comporte pas spécialement de fonctionalités élaborées.

[5] WEBCRAWLER fait remarquablement exception. Cf. [Pinkerton] et notre commentaire en Webliographie.

moyen grossier de capter des phénomènes linguistiques et textuels), que d'identifier les propriétés que ces formules veulent retranscrire.

De même, il n'est pas dans le champ de notre étude de repérer les astuces d'implémentation, qui sont pour beaucoup dans l'efficacité du traitement (en termes de rapidité et éventuellement de stockage)[6]. Ce n'est d'ailleurs plus là un goulet d'étranglement : les performances des matériels, déjà impressionnantes, continuent à s'accroître rapidement, et ce n'est plus sur ce plan que pourraient se distinguer les recherches les plus efficaces[7].

[6] Nous citons cependant en Bibliographie-Webliographie [Manber] qui s'intéresse à ces questions.

[7] « On the surface, it would seem that each search engine should be able to find the same sites based on a given query, so you might think that the main difference would be speed. However, our tests found the opposite to be true. We found the differences in performance to be negligible. » [Venditto]
« It is impossible to make a meaningful assessment of the speed of searching. All the search tools give a rapid response under optimum conditions, but network congestion can make the latter seem infrequent. [...] When I started this brief investigation [about World Wide Web searching tools] I imagined that retrieval performance would be most important, but record structure and search techniques look to be of greater significance. » [Winship]

2. Résultat de l'enquête : récapitulatif systématique des paramètres

Ce chapitre constitue le cœur de la note, il fait la synthèse de l'ensemble des idées recensées. Nous avons organisé ces idées selon un plan lui-même porteur de sens, puisqu'il met en évidence et structure les unités de décision ou de traitement des moteurs de recherche.

Les parties **I.**, **II.**, **III.**, **IV.** s'enchaînent de façon chronologique (chacune suppose réalisées les opérations décrites par les précédentes). La première partie s'intéresse à la constitution de la base (fonds documentaire sur lequel s'effectue les recherches), elle peut ne concerner que des traitements en mode « batch » ; elle s'oppose en cela aux parties suivantes (**II.**, **III.**, **IV.**), qui s'articulent autour du traitement d'une requête, sur un mode interactif / temps réel.

Les parties se divisent ensuite en chapitres (**A.**, **B.**, **C.**, etc.) qui se focalisent sur une étape du traitement -sans nécessairement d'organisation chronologique cette fois-ci.

La description se poursuit en subdivisions successives. A chaque nœud, les différentes possibilités ne sont pas exclusives. L'arbre obtenu, une fois déployé, fonctionne un peu comme un questionnaire, une grille de comparaison, sur laquelle chaque ligne représente une fonctionnalité pouvant être réalisée ou non, présente ou non, dans un outil de recherche.

Ainsi sont récapitulés, de façon systématique et organisée, l'ensemble des caractéristiques et fonctionalités des services actuels de recherche automatisée sur WWW à base d'index.

I. Base de données
- **I.A.** Constitution de la base
  - **I.A.1.** Nature des données accessibles
    - **I.A.1.a.** Sources
      - HTML
      - FTP
      - Usenet News
      - Gopher
      - etc.
    - **I.A.1.b.** Média (technologie de lecture spécifique)
      - images GIF
      - Java
      - VRML
      - Acrobat
      - etc.
  - **I.A.2.** Visée exhaustive vs. sélective[8]
    - **I.A.2.a.** Volume (capacité)
    - **I.A.2.b.** Couverture
      - Domaine
        - général
        - géographique
        - thématique
        - chronologique (nouveautés)
      - Stratégie de parcours et de recueil de l'information [cf. **V.A.2.a.**]
        - visite des liens en « largeur d'abord » (diversité des serveurs)
        - visite des liens en « profondeur d'abord » (documents plus spécialisés)
    - **I.A.2.c.** Possibilité de filtrage sélectif paramétrable
  - **I.A.3.** Mise à jour
    - **I.A.3.a.** Fréquence
      - Entrée de nouveautés
      - Corrections, modifications

[8] Le réseau est tellement immense qu'il faut bien se donner une politique de parcours. Par exemple celle-ci :
« To compile its database, WEBCRAWLER surveys the entire Web, evaluating the popularity of each site and storing only the contents of pages that seem well-traveled or that fill in gaps in the existing database. » [Venditto]

I.A.3.b. Mode (portée sur l'organisation de la base)
- Cumulative, incrémentale
  (prise en compte seulement de ce qui change : impact local)
- Générale
  (réorganisation de toute la base, impact global des changements)

I.A.4. Entretien[9]
I.A.4.a. Elimination des liens périmés (inactifs - « dead links »)
I.A.4.b. Dédoublonnage (groupement des références multiples pour un même document)
- Versions successives
- Copies sur différents serveurs
- Variantes d'adresse du serveur (aliases)

I.B. Descriptif intrinsèque du document
(Nous ne décrivons pas le cas de données standardisées,
où chaque auteur de document remplit spécialement pour la base un formulaire descriptif fourni,
cf. par exemple ALIWEB.)

I.B.1. Données factuelles recueillies
I.B.1.a. Adresse
- Site, localisation du serveur (telle université, tel centre de recherche[10]...)
- Zone géographique (pays)
- Adresse(s) alternative(s) : autres versions du même document, doublons

I.B.1.b. Date, données temporelles
- Auteur
  date d'édition ou de dernière modification
  fréquence habituelle de révision
- Base du moteur de recherche
  date de repérage/enregistrement par le moteur pour indexation
  date d'intégration à la base (dans l'espace de recherche)
  fréquence habituelle de visite pour mise à jour
- Lecteurs
  nombre de visites, fréquence d'accès à la page

I.B.1.c. Taille du fichier
I.B.1.d. Liens hypertexte
- Documents citant
  nombre de liens pointant sur le site[11]
  référence (adresse) de documents citant
- Documents cités
  référence (adresse) des documents cités

I.B.1.e. Langue de rédaction

I.B.2. Données textuelles recueillies
I.B.2.a. Titre
- Nom général du document
- Titre du document (titre(s) de premier niveau)
- Tous les niveaux de titre

I.B.2.b. Méta-informations [12] (fournies par l'auteur)
- Descriptif (champ spécifique, renseigné librement et manuellement, par l'auteur)
- Mots-clés (déclarés comme tels par l'auteur dans une zone spécifique)

I.B.2.c. Texte intégral
(vu comme une suite de chaînes de caractères :
constitution d'un fichier inverse (index))

---

[9] Les opérations d'élimination des liens inactifs et des doublons sont souvent plutôt proposées en tant que post-traitement : avant de présenter les résultats d'une requête, certains moteurs « nettoient » ainsi la liste des documents à afficher. Avec cette logique, la vérification est alors entièrement réeffectuée pour les sélections de chaque requête.

[10] Les champs en lesquels se décompose l'adresse peuvent ensuite être maniés comme des données textuelles (notamment, recherche sur des sous-chaînes de caractères comme « .edu », « to constrain your search to educational institutions. », comme le propose HOTBOT en exemple).

[11] Utilisé comme indice de popularité, de qualité, d'intérêt.
Ce serait même un des indices les plus fiables, permettant de contrecarrer les manœuvres de certains documents pour apparaître abusivement comme pertinents (cf. **V.A.4.a.**) : « one of the best ways to avoid the spamming problem without overt penalties might be to give more weight to a page's popularity [...]. A page with lots of links to it is in effect « ranked » by those across the web. Good pages get lots of links ; bad pages don't. » [Sullivan]

[12] Les balises SGML qui les repèrent sont des « méta-tags ». Le texte donné entre ces balises décrit le document pour qu'il puisse être catalogué correctement ; il n'est pas affiché quand on visualise le document lui-même.

I.C. Traitement et codage des données textuelles (réduction)
- **I.C.1.** Sélection d'extraits représentatifs (repérés automatiquement)
  - **I.C.1.a.** Vocabulaire / concepts
    - Les *n* mots les plus fréquents, avec leur fréquence (par exemple *n*=100)
    - Zone d'ancrage textuel des liens hypertextes[13]
  - **I.C.1.b.** Résumé automatique (par extraction)
    - Début du texte
      - les *n* premières lignes[14] (*n*=20, ou *n*=min(20,20%) par exemple)
      - les *n* premières phrases (repérées par la ponctuation)
    - Début (et fin) de « paragraphes-clé »
    - Sélection de phrases comprenant la plus forte densité de termes saillants[15]
- **I.C.2.** Elimination de « mots-vides »
  - **I.C.2.a.** Critères formels (sur la chaîne de caractères)
    - Mots de 1 ou 2 lettres
    - Chaînes de caractères comprenant autre chose que des lettres[16]
  - **I.C.2.b.** Critères linguistiques
    - Syncatégorématiques (mots grammaticaux) -donnés dans une liste.
  - **I.C.2.c.** Critères documentaires
    - Mots liés au type de document (par exemple ici « WWW »)
    - Mots liés au domaine de la base
- **I.C.3.** Regroupement
  - **I.C.3.a.** Typographique
    - Distinction de casse
      - minuscules et majuscules sont confondues
      - minuscules et majuscules sont distinguées[17]
        - pour toute chaîne de caractères
          - possible (option explicite)
            - pour l'ensemble de la requête
            - pour chaque terme de la requête
          - imposé
        - pour les chaînes mêlant les deux casses
        - pour les chaînes tout en majuscules
        - pour les chaînes où il y a au moins une majuscule pas en première position
    - Accentuation (et autres diacritiques)
      - les accents sont neutralisés
      - les accents mis à un terme de requête sont pris en compte
    - Ponctuation
      - tous les signes sont confondus et équivalents au blanc (notamment le tiret)
      - les points dans les sigles en majuscules sont retirés, et les lettres recollées
  - **I.C.3.b.** Flexionnel[18]
    (lemmatisation, en général sommaire, au moyen d'une liste de terminaisons standard -pluriel, conjugaison-)
    - Identité (on néglige les variantes en ramenant tout au type reconnu)
    - Equivalence (on garde en mémoire la forme fléchie)

[13] Ce sont les mots sur lesquels on peut « cliquer ». La présentation HTML standard les affiche en bleu souligné.
[14] « Quoique la documentation [des applications utilisant cette tactique] ne fournisse aucune explication concernant la définition d'une ligne (car il n'y a pas de notion de ligne en HTML). » note [Plourde].
[15] Là encore, évidemment, peu d'explications. Les indices de saillance les plus connus sont définis à partir de la fréquence et/ou du caractère discriminant (en étudiant la distribution du vocabulaire sur le corpus).
[16] Il faut bien peser l'impact de cette décision : l'éviction des chiffres nuit à la reconnaissance des dates ; celle des symboles, comme « + », à celle de sigles comme « C++ ».
[17] La distinction minuscule vs. majuscule est utile pour beaucoup de noms commerciaux (entreprises, produits), volontiers métaphoriques. Les exemples donnés à ce propos dans [Barlow] sont parlants : « many words in English are used both as proper and common nouns -Bill, bill, Gates, gates, Oracle, oracle, Lotus, lotus, Digital, digital- the list is endless. »
[18] La réduction flexionnelle rapporte le nom pluriel à sa forme au singulier, l'adjectif à sa forme masculin singulier, le verbe conjugué à son infinitif. Le stade suivant est la réduction dérivationnelle, qui regroupe les mots de la même racine (« mots de la même famille »). La réduction dérivationnelle est bien sûr encore plus difficilement automatisable, et nous ne l'avons pas trouvée implémentée ; cependant, certains procédés de troncature (cf. **II.A.1.a.**) permettent de jouer sur les variations autour d'une racine lexicale.

I.C.3.c. « Conceptuel »
- Construction automatique de classes d'équivalence de termes
association des termes qui apparaissent dans les mêmes contextes
(corrélation des termes par rapport à leur distribution
sur l'ensemble de documents)

I.C.4. Informations quantitatives (pour chaque mot)
I.C.4.a. Nombre d'occurrences dans le document
I.C.4.b. Position(s) dans le document

II. Lancement d'une recherche : Forme de la requête, interface d'interrogation
II.A. Langage d'interrogation[19]
II.A.1. Lexique
II.A.1.a. Vocabulaire non contrôlé (pas de dictionnaire des termes autorisés)
- Chaînes de caractères (vision « informatique »)
frontières
sous-chaînes bornées
(on recherche sur les mots limités par les blancs)[20]
caractère joker (masque) -éventuellement d'usage restreint :
puissance
remplace un caractère
remplace une sous-chaîne limitée[21]
position
en dernière position (troncature à droite)
en première position (troncature à gauche)
après au moins *n* caractères (par exemple *n*=3)
expressions régulières
recherche modulo une tolérance orthographique[22]
limitation du nombre d'erreurs (seuil)[23]
calcul d'une distance entre chaînes de caractères

[19] La « langue naturelle » est ici considérée comme un langage d'interrogation, au même titre que d'autres langages dédiés à la formulation de requêtes (point de vue fonctionnel).
[20] L'effet des frontières est sensible soit dans les recherches les plus simples (la requête prend la forme d'une suite de termes, non séparés par des opérateurs), soit dans les recherches élaborées où l'on peut utiliser des caractères « jokers » (il est alors d'usage que ceux-ci ne recouvrent pas de blanc-espacement). Ainsi :
« Regular expressions will generally not cross word boundaries (because only words are stored in the index). So, for example, « lin.*ing » will find « linking » or « flinching », but not « linear programming ». » [HARVEST BROKER]
[21] Par exemple, ALTAVISTA propose la notation « * » avec la portée suivante :
« The *-notation cannot be used witout restriction. To make such queries computationally feasible, ALTAVISTA requires that the * be used only after at least three letters. The *-notation will match from zero up to five additional letters in lower-case only. Capital letters and digits will not therefore be matched. »
[22] Selon les contextes applicatifs, on peut prévoir différents types d'erreurs, qui seront corrigées par des stratégies différentes ; par exemple : faute d'orthographe à proprement parler (mauvaise retranscription d'une information phonétique) ; faute de frappe (inversion, omission, redoublement ou ajout de caractère, influence du voisinage des touches sur le clavier) ; mauvaise reconnaissance optique du caractère (dans le cas d'une acquisition par OCR). Voir notamment [Manber].
[23] Option originale, proposée par HARVEST BROKER :
« Number of errors allowed : Glimpse [HARVEST BROKER's search engine] allows the search to contain a number of errors. An error is either a deletion, insertion, or substitution of a single character. The Best Match option will find the match(es) with the least number of errors. The default is 0 (zero) errors. »

- Descripteurs libres (vision « linguistique »)
  - recherche modulo les variantes de réduction (lemmatisation, mots-vides)
    - impossible (l'utilisateur doit décrire toutes les variantes s'il les veut)
    - possible
      - terme à terme
      - globalement pour l'ensemble des termes
    - imposé (l'information des formes originales des occurrences n'a pas été gardée pour le traitement)
  - recherche modulo d'autres régularités morphologiques particulières
    - nom propres de personnes[24]
      - ellipse du nom ou du prénom
      - utilisation d'initiales (éventuellement celle d'un second prénom)
      - inversion de l'ordre (dans une bibliographie par exemple)
  - recherche modulo des associations conceptuelles
    - impossible (l'utilisateur doit penser aux divers synonymes)
    - possible
      - pour chaque terme indépendamment
      - globalement pour l'ensemble des termes
    - imposé (les termes originaux ont été remplacés par les concepts)

II.A.1.b. Terminologie fermée
(correspond plutôt à l'approche complémentaire, par catalogue/classement, où les documents sont repérés par rapport à un ensemble de rubriques déterminées)

II.A.2. Syntaxe

II.A.2.a. Séparateurs définissant les termes (atomes) de recherche[25]

- Le caractère blanc d'espacement définit les unités : recherche par « mot »
- Possibilité de recherche par groupe de mots ou syntagme (« phrase », en anglais[26])
  - Elimination des mots-vides
    - impossible (l'utilisateur doit prévoir par exemple des variantes de prépositions)
    - possible
      - partielle (la position et le nombre des mots-vides sont conservés : 1 mot-vide = 1 joker)
      - totale
    - imposée[27]
      - partielle (la position et le nombre des mots-vides sont conservés : 1 mot-vide = 1 joker)
      - totale

---

[24] « If you capitalize adjacent names [without using commas to separate them], INFOSEEK GUIDE treats the words as a single name. » [INFOSEEK GUIDE]
HOTBOT utilise une grammaire un peu plus élaborée [nous avons juste changé le nom pris en exemple pour que l'explication soit plus claire] :
« The « Search the Web for a person » search option uses a limited proximity search to find the most possible matches. For example, a three-word name like JOHN A. SMITH, will match the following phrases :
JOHN A SMITH
JOHN SMITH
SMITH, JOHN
A two-word name like JOHN SMITH will match the following phrases :
JOHN SMITH
SMITH, JOHN »

[25] Nous utiliserons par la suite le mot « terme » dans ce sens, à savoir pour désigner les mots, ou les groupes de mots pris comme une seule expression, sur lesquels s'effectuent les recherches élémentaires, et qui sont ensuite éventuellement combinés dans une équation booléenne ou pseudo-booléenne. Ce sont bien des « atomes » de recherche, au sens étymologique, puisqu'on ne les décompose pas (on ne s'intéresse pas aux documents qui ne contiennent qu'une fraction ou qu'un fractionnement de l'atome).

[26] Attention au faux-ami : le mot anglais « phrase » désigne une expression, un mot composé, un syntagme (groupe nominal), bref un petit groupe de mots syntaxiquement étroitement liés dans une phrase. Souvenez-vous que la phrase se traduit, elle, par « sentence ».

[27] Cela est dans certains cas pénalisant. En effet, une suite de mots très communs peut être, elle, très spécifique et précise, et constituer une requête valable ; par exemple, une recherche de « To be or not to be » serait refusée par un moteur éliminant d'emblée les « mots vides ».

II.A.2.b. Qualification des termes de la requête par l'utilisateur
- Pondération numérique libre
- Modes de valorisation (importance accordée à un terme)
  - indicateur de présence obligatoire (souvent noté « + »)
  - indicateur d'importance, de rôle prépondérant[28]
- Modes de limitation d'influence
  - indicateur de caractère secondaire (rare ; noté « - » quand ce symbole ne traduit pas l'opérateur d'exclusion)
  - rôle simplement dans la sélection
  - rôle simplement dans le classement pour la lecture des résultats

II.A.2.c. Qualification des termes du document par les balises
- Présence dans une zone textuelle
  - nom du document
  - titre
  - descriptif
  - texte
  - ancrage textuel
    - d'un lien
    - d'une applet Java[29]
- Présence dans une zone descriptive
  - adresse (URL)
    - du document
    - d'un document cité[30]
      - lien externe (autre serveur)
      - lien interne (vers une image par exemple)
  - rubriques des Usenet News[31]

II.A.2.d. Longueur de la requête
- Limitation du nombre maximum de termes dans la requête
- Limitation sur le nombre de termes à considérer simultanément (cas de requêtes non parenthésées et décomposables avec une combinatoire de sous-parties)
- Limitation du nombre de documents concernés par un terme[32]

II.A.2.e. Logique booléenne
- Articulation
  - application d'un seul opérateur à l'ensemble des termes de la requête : ANY, ALL
  - combinaison complexe d'opérateurs (parenthésage)

[28] La seule proposition que nous avons relevée dans ce sens, c'est celle d'un système, qui conseille, pour mettre en valeur un terme de la requête, de le répéter ! Voir aussi la discussion sur pondération quantitative et pondération qualitative, un peu plus loin dans ce document.
[29] Par exemple un texte qui défile, ou qui clignote.
[30] L'utilisation de l'information sur les liens vers des documents extérieurs serait à rapprocher de l'analyse des (co)citations, traditionnelle en bibliométrie.
[31] Exemples de rubriques des Usenet News pouvant être sigificatives ici : auteur-expéditeur, newsgroup destinataire, sujet, résumé, mots-clés.
[32] C'est-à-dire qu'une combinaison de termes généraux, même très précise, sera rejetée. Ces limitations viennent en fait des trop grandes ressources de mémoire et de calcul qu'il faut mobiliser pour ce genre de requête (combinatoire élevée des possibilités à tester).

- Opérateurs
  - conjonction, disjonction : AND, OR
  - exclusion (négation) : NOT (parfois noté « - »)
  - proximité : NEAR[33]
    - paramétrable :
      - distance en caractères (par exemple *n*=80)
      - distance en mots[34] (par exemple *n*=10 ou 100)
      - appartenance à un même champ (titre, descriptif, etc.)
      - appartenance à un même paragraphe
      - appartenance à une même phrase (ponctuation)
  - ordre : FOLLOWED BY (ici : adjacence)

II.B. « QBE » = query by example

II.B.1. A partir de documents de la base

II.B.1.a. Feed-back à partir de la liste des résultats
(conservation du point de vue de la requête initiale)[35]

- Utilisation d'indications thématiques préalablement associées au document
- Utilisation des liens, en fonction de leur ancrage textuel, pour ajouter certains documents cités à l'espace de recherche
- Utilisation du texte[36]
  - contexte des mots ayant contribué à la sélection du document

II.B.1.b. Rebondissement à partir de la liste des résultats
(lancement d'une nouvelle requête à partir d'un document trouvé)

- Utilisation du texte (cf. **II.B.2.**)
  - données textuelles brèves (cf. **I.B.2.a.**, **I.B.2.b.**, **I.C.1.**)

II.B.2. A partir d'un texte tapé ou « collé » comme requête
(ce texte n'est pas nécessairement référencé dans la base, on dispose donc a priori de moins de données)

II.B.2.a. Traitement identique à celui d'une suite de descripteurs libres

II.B.2.b. Traitement spécifique (propriétés textuelles)

III. Traitement d'une recherche : Calcul / évaluation de la pertinence

III.A. Mode de prise en compte des opérateurs

III.A.1. Booléen : solutions exactes (« exact match »)

III.A.2. Booléen « flou » : solutions les meilleures (« best match »)

III.A.2.a. Une conjonction dont les termes ne sont pas tous présents est pénalisée, mais pas nulle

III.A.2.b. Une disjonction (non exclusive) dont plusieurs termes sont retrouvés est avantagée.

III.B. Prise en compte de la distribution des occurrences des termes de recherche

III.B.1. Un document est d'autant plus pertinent pour un terme si ce terme :

III.B.1.a. (présence)

- apparaît dans peu de documents de la base (il est « discriminant »)

---

[33] NEAR est toujours implicitement proposé comme un opérateur binaire. On pourrait concevoir un opérateur *n*-aire, tel que par exemple tous les termes arguments auxquels il s'applique soient dans un voisinage d'extension déterminée (un même paragraphe, une fenêtre de 50 mots,...).
Remarquons enfin que, pour le cas de l'adjacence ordonnée, on a classiquement un opérateur *n*-aire, à savoir les guillemets qui encadrent une expression (chaîne de mots à chercher). Autrement dit :

« diffusion automatique ciblée » (notation par guillemets)

équivaut à

diffusion FOLLOWED BY automatique FOLLOWED BY ciblée (notation par opérateur binaire)

On voit clairement que la seconde forme est moins élégante, ce qui pourrait expliquer qu'un opérateur de type FOLLOWED BY (pour signifier l'adjacence) soit effectivement rarement proposé.

[34] La distance en mots sert souvent à gérer l'adjacence (distance nulle) et l'appartenance à un même syntagme (distance de l'ordre de 0, 1, 2 mots). L'appartenance à un même syntagme requerrait de très grosses ressources, en termes d'outils linguistiques, pour être traitée proprement (avec un point de vue syntaxique).

[35] Nous n'avons trouvé nulle part la possibilité d'indiquer simultanément plusieurs documents : la requête est relancée à partir d'un seul document.

[36] La richesse lexicale des textes aide à pallier l'absence de vocabulaire contrôlé ; en effet, consulter un certain nombre de documents sur le sujet recherché inspire pour enrichir et préciser la requête initiale avec d'autres termes voisins, auxquels on n'avait d'abord pas pensé. L'enjeu est de retrouver cette stratégie dans un traitement automatisé.

III.B.1.b. (fréquence)
- a un grand nombre d'occurrences dans le document
- a un grand nombre d'occurrences par rapport à la longueur totale du document
- n'est pas un hapax dans ce document (une seule occurrence)

III.B.1.c. (position)
- est proche du début du texte[37]
- est dans une zone significative (par exemple, titre > mots-clés > texte intégral)

III.B.1.d. (normalisation)
- est trouvé tel quel (plutôt qu'une de ses variantes)

III.B.2. Un document est d'autant plus pertinent vis-à-vis d'un ensemble de termes, que :

III.B.2.a. (co-présence)
- il comporte un plus grand nombre de ces termes

III.B.2.b. (positions relatives)
- les termes ont des occurrences proches les uns des autres
- ces termes sont retrouvés dans l'ordre de la requête

IV. Des résultats à la réponse

IV.A. Affichage des résultats et aide à l'interprétation

IV.A.1. Données caractérisant la requête (appréciation du volume et de la répartition des documents concernés)[38]

IV.A.1.a. Nombre de documents trouvés en réponse
- Absolu
- Relatif (rapporté au volume total de la base)
- Nombre de documents où apparaissent *n* termes (*n*=1,2,...jusqu'au nombre total de termes soumis)

IV.A.1.b. Nombre de serveurs comprenant des documents réponse
- Absolu
- Relatif (rapporté au nombre total de serveurs répertoriés dans la base)

IV.A.1.c. Données associées à chaque terme de la requête
- Nombre d'occurrences dans l'ensemble de la base
- Nombre de documents où apparaît le terme
- Données détaillées sur les termes indiqués comme étant « à l'essai »

IV.A.2. Données intrinsèques au document (pour chaque document sélectionné)[39]

IV.A.2.a. Lien actif

IV.A.2.b. Données factuelles[40] (cf. **I.B.1.**)

IV.A.2.c. Données textuelles brèves (cf. **I.B.2.a.**, **I.B.2.b.**, **I.C.1.**)

IV.A.2.d. Données et index utilisés pour caractériser le document lors du calcul de la pertinence

IV.A.2.e. Appréciation[41] (cela déborde le cadre d'une approche tout automatique)

---

[37] Voici comment nous pourrions interpréter le caractère significatif accordé au début du texte : cette position peut traduire :
- l'appartenance à un niveau élevé dans la structure hiérarchique du texte (titre, intertitre), niveau où l'on trouve des expressions synthétiques et résumantes de l'ensemble du texte ;
- le fait qu'il ne s'agisse pas d'un sous-thème, abordé dans un développement ;
- la mise en valeur par le rédacteur, notamment pour des formes de type journalistique, pour lesquelles le début du texte doit concentrer l'essentiel de l'information.

[38] Dans un système couvrant une base moins volumineuse, on pourrait imaginer que des indications sur l'impact de la requête soient données au fur et à mesure de l'élaboration de la requête. On verrait par exemple au moment de l'ajout d'un terme s'il contraint trop, ou élargit trop, l'ensemble des résultats.

[39] Certains services fourniront des informations de classement, une analyse critique,... C'est une valeur ajoutée offerte, mais que nous ne considérons pas ici, car nous voulons rester dans le cadre de systèmes où le traitement des documents est automatisé.

[40] Notamment la donnée explicite de l'adresse (URL), pouvant ainsi être notée au cas où la connexion est momentanément impossible ou pour un usage ultérieur.

[41] Par exemple, les « étoiles » (de une à quatre) de MAGELLAN cherchent à caractériser la qualité du document pour lui-même. Il ne faut pas confondre cet indicateur avec une évaluation de la pertinence, elle *relative* à une requête donnée.
Les critères que se fixe MAGELLAN sont intéressants :
« - depth : is it comprehensive and up-to-date ?
- ease of exploration : is it well-organized and easy to navigate ?
- net appeal : is it innovative ? Does it appeal to the eye or the ear ? Is it funny ? Is it hot, hip, or cool ? Is it thought-provoking ? Does it offer new technology or a new way of using technology ? »

IV.A.3. Informations sur le rapprochement
IV.A.3.a. Données associées à chaque terme commun à la requête et au document
- Variantes trouvées dans le document
  (variantes avant lemmatisation par exemple)
- Nombre d'occurrences dans le document
- Pondération utilisée
- Contexte(s) d'occurrence de chaque mot dans le document
  (« KWIC » = key-word in context)
- surlignage[42] des occurrences dans le texte intégral
- extraction des phrases ou des passages les plus significatifs
  (ayant le plus contribué à la sélection du document par cette requête)[43]

IV.A.3.b. Mesure de pertinence
- Valeur numérique normée (par exemple comprise entre 0 et 100)
  absolue (la valeur maximale correspond au critère le plus fort
  de proximité, quand les documents apparaissent identiques)
  relative à l'ensemble des résultats
  la valeur maximale est celle du meilleur rapprochement obtenu
  la valeur maximale est fonction de la force / fiabilité
  du meilleur rapprochement obtenu[44]
- Icône
  (traduisant de façon synthétique et visuelle le degré de pertinence calculé)

IV.A.4. Organisation de l'ensemble des rapprochements
IV.A.4.a. Liste avec numérotation explicite[45]
IV.A.4.b. Ordre
- Valeur de pertinence décroissante
- Nombre de liens décroissant
  nombre de documents pointant sur le document avec un lien hypertextuel
  (interprété comme un indice de popularité, de qualité, d'intérêt)
- Chronologie (mise en valeur des nouveautés)
- Taille du document
- Adresse (URL), par nombre de documents sélectionnés décroissants
  (mise en évidence de pôles de compétences)

IV.A.4.c. Classement par rubrique
- Nature du document (cf. **I.A.1.a.**)
- Adresse (URL), par ordre alphabétique
- Domaine du document
  combinaison du moteur de recherche avec un catalogue thématique
  utilisation d'une vision par « concept » pour avoir une vision synthétique[46]
- Présence de mots indiqués par l'utilisateur

IV.A.4.d. Organisation selon les liens éventuels entre les documents trouvés[47]

Si cela peut apporter des éléments pour interpréter et utiliser les résultats de la recherche, cela reste un point de vue particulier, par ailleurs difficilement automatisable. Mais la grille de notation méritait d'être citée.

[42] Surlignage par surbrillance, clignotement,...

[43] EXCITE propose ainsi un résumé automatique (« Automatic abstracting ») contextuel :
« Another capability enabled by ICE searching is the ability to examine selected documents and then abstract those sentences which are judged relevant to the concept. Thus the user can create automatic abstracts of the documents on their matching list, to speed the process of evaluation. »

[44] Par exemple, si la pertinence semble faible même pour le meilleur document, alors la valeur maximale n'est pas 100, mais 99. C'est ce que fait HOTBOT :
« Once the set of matching documents has been identified, the scores are normalized so that no document scores over 100%. If all the documents are poor matches, the best of the poor matches is given a score of 99%. » [HOTBOT]
C'est clairement une convention, qui doit être explicitée pour une interprétation juste des résultats.

[45] Numéroter les résultats n'a peut-être l'air de rien ! Et pourtant, c'est très appréciable pour se repérer quand on parcourt la liste des résultats.

[46] EXCITE propose un « Automatic subject grouping ».

[47] L'idée de faire apparaître, dans la présentation des résultats, les liens entre les documents présentés, était exposée dans un article prospectif ; nous n'avons vu nul part cette fonctionnalité implémentée. Il est vrai qu'elle soulève un certain nombre de difficultés : une simple organisation hiérarchique (père = document citant, fils = document cité), comme envisagée par l'article, ne serait évidemment pas suffisante (des documents peuvent se citer réciproquement, un document peut être cité par plusieurs autres, etc.) ; et la représentation de tous les liens risque de mener à une visualisation peu lisible, en raison de la surcharge et de l'intrication des liens. Une solution élégante (expérimentée dans le cadre de la Bibliothèque Electronique, projet EDF / DER / IPN / SID) est de ne rendre visible les liens que sur demande, pour un document pointé par l'utilisateur.

IV.B. Maniabilité (notamment dans la perspective d'affinement de la recherche par itérations)
IV.B.1. Volume transmis
IV.B.1.a. Rapidité du calcul et de la transmission des résultats[48]
- Pas de dégradation des performances avec la montée en volume (« scalability »)
- Indicateur de la charge du serveur

IV.B.1.b. Gestion de la surcharge
- Surcharge pour le traitement ou le transfert des fichiers
  - présentation des résultats par tranches (transmission de fichiers de taille raisonnable)
  - possibilité d'interruption de la recherche[49]
  - option avec/sans images[50]
- Surcharge cognitive
  - présentation ordonnée, avec en premier les documents a priori les plus utiles
  - seuillage
    - nombre de documents proposés
    - valeur suffisante de l'indicateur de pertinence[51]
    - filtre chronologique
      - absolu (repère ponctuel) : avant / après telle date
      - relatif (durée) : depuis *n* semaines / mois / années
  - paramétrage du niveau de détail des sorties[52]
    - informations développées
    - informations standard (par défaut)
    - informations abrégées
    - simple mention des documents

IV.B.2. Clarté, simplicité d'emploi
IV.B.2.a. Convivialité, ergonomie
- Abord simple pour les novices (intuitif)
- Simplicité d'utilisation
  - menus (qui détaillent l'éventail des possibles et soulagent la mémoire)
  - pas d'enchaînement de commandes
- Possibilité de formulations expertes plus puissantes (pas de nivellement par le bas)
- Efficacité rapide
  - efficacité de la recherche standard, par défaut
  - efficacité de la recherche en un coup (quand on dispose de peu de temps pour retravailler sa requête)

---

[48] La rapidité de traitement ne serait plus un facteur discriminant : les moteurs de recherche connus ont tous de bonnes performances sur ce point.

[49] La possibilité d'interruption est en faite offerte par le browser (Netscape, Mosaic, etc.)

[50] Certains récepteurs préfèrent une version sans images, pour des questions de lenteur et/ou de coût. (Les images sont en effet des données beaucoup plus volumineuses que le texte.)

[51] C'est ce à quoi correspondent, de façon fort peu claire (et là les comparatifs sont unanimes), les rubriques de LYCOS : « Loose match / Fair match / Good match / Close match / Strong match ». Tel que cela est présenté, on croit qu'il s'agit d'un ajustement plus ou moins lâche sur la forme des termes. Mais en fouillant bien la documentation associée à l'outil (à la fin des FAQ, après tout un passage plus spécifique et technique qui aurait arrêté plus d'un lecteur), nous avions fini par trouver que chacune de ces expressions correspondait à un seuillage sur la valeur de pertinence calculée : Loose match prend tous les documents dont le coefficient de pertinence est au moins égal à 0.1, Fair match seuille à 0.3, Good à 0.5, Close à 0.7, et Strong à 0.9.
Mais peut-être que cette précision dans l'aide en ligne ne fut que récente et passagère (elle a disparu avec les FAQ), car même nos spécialistes avouent honnêtement leur perplexité ou leur désarroi :
« You're [...] able to set controls for the degree of relevancy desired in search results. Although LYCOS gives you the choice of how strictly the search terms should be enforced, it does not provide much documentation on how you can use the controls to improve the quality of your search. » [Venditto] (et pour cause ! Cela ne modifie en rien les éléments de la liste, cela ne fait que couper celle-ci un peu plus haut ou un peu plus bas...)
« Search refining : [...] the rather unclear « loose match, fair match, good match, strong match, or close match. » (Huh ??) » [Barlow]
« Il est possible aussi de spécifier le degré de précision des termes recherchés (loose, fair, good, close, strong match). Nous devons admettre cependant que nous n'avons pas saisi toute la subtilité entre ces paramètres ! » [Plourde]

[52] Plutôt que des niveaux de détail prédéfinis, on pourrait concevoir que l'utilisateur puisse cocher, dans un ensemble d'informations descriptives disponibles proposé, celles qui l'intéressent pour qualifier les résultats de sa requête.

IV.B.2.b. Documentation technique claire et complète (cf. fichiers « Help »)[53]
- Contextualisation de l'aide[54]
- Manuel d'utilisation
  - description de la syntaxe
  - description des paramètres et options de recherche
    - à partir d'un schéma fléché de l'interface (écran)
  - exemples
    - montrant la forme des requêtes
    - montrant la forme des résultats
    - expliquant l'interprétation d'un résultat
    - justifiant l'utilité des fonctionnalités (cas d'application utile)
  - mémento récapitulatif des fonctionalités de recherche[55]
- Conseils (notamment pour la (re)formulation des requêtes)
  - pour gagner en précision
    - utiliser avec à propos les contraintes sur les accents, la casse
    - utiliser des expressions plutôt que de mots isolés
    - bien préciser sa pensée en explicitant tous les termes[56]
    - éviter les termes trop généraux
    - utiliser les options de recherche,
      notamment les opérateurs d'exclusion[57] (cf. **II.A.2.e.**),
      la recherche par syntagme (termes composés) (cf. **II.A.2.a.**)
    - trouver des idées de termes appropriés et complémentaires
      dans les documents pertinents déjà trouvés

[53] La qualité de la documentation du système est essentielle, elle conditionne sa bonne utilisation. Elle a bien des rôles : indiquer la mise en forme d'une requête, guider dans l'utilisation des options, expliquer comment interpréter les résultats retournés, donner des orientations pour améliorer ou affiner les résultats : c'est dire qu'elle est importante à tous les stades de la recherche.
« La documentation pour les services de repérage aide les utilisateurs à atteindre deux objectifs. Le premier est d'évaluer la pertinence de la base, c'est-à-dire sa nature (objets répertoriés), ses objectifs, son autorité, etc. Le second est la maîtrise et l'utilisation efficace des services de repérage et la vérification du comportement de ces services (obtient-on les résultats escomptés ?). » [Plourde]
« La documentation pour les services de présentation des résultats assiste l'utilisateur à interpréter correctement les résultats. Par exemple, [...] plusieurs outils de recherche offrent un classement des résultats obtenus selon leur pondération par rapport à la requête de recherche. Malheureusement peu d'entre eux expliquent le calcul de cette pondération. » [Plourde]
« If you understand how search engines organize information and run queries, you can maximize your chances of getting hits on URLs that matter. » [Barlow]

[54] Par exemple, l'information sur la mise en forme des requêtes élaborées est la première accessible depuis la page de soumission des requêtes élaborées, et n'est qu'indirectement acessible depuis la page de soumission des requêtes simples.

[55] INFOSEEK GUIDE fournit ainsi une page qui résume la syntaxe d'interrogation (« Quick référence to syntax ») sous la forme (claire) d'un petit tableau.

[56] Conseil répété s'il en est, car les habitudes sont de ne lancer qu'un ou deux mots-clés !
[Pinkerton] a le mérite de faire constater explicitement ce phénomène, et de vouloir activement le prendre en compte :
« Another factor limiting the precision of queries is that users do not submit well-focused queries. In general, queries get more precise as more words are added to them. Unfortunately, the average number of words in a query submitted to WEBCRAWLER is 1.5, barely enough to narrow in on a precise set of documents. I am currently investigating new ways to refine general searches and to give users the ability to issue more precise queries. » [Pinkerton]

[57] A notre avis, ce conseil doit être reçu avec beaucoup de réserves : les exclusions sont dangereuses, car souvent trop puissantes (et de portée insoupçonnée).

pour réduire le silence[58]
s'assurer qu'il n'y a pas de faute de frappe (!)
vérifier l'orthographe des termes (notamment les variantes entre anglais britannique et américain)[59]
vérifier la syntaxe de la requête[60]
ajouter des variantes flexionnelles (pluriel, conjugaison,...)
ajouter des variantes dérivationnelles (mots de la même famille)
mettre l'abréviation et sa forme développée pour les sigles, acronymes, etc.
ajouter des synonymes (éventuellement en s'inspirant de documents déjà retrouvés)
lancer la requête sur une base moins sélective

- Réponse aux principales questions posées sur le système (« FAQ » = frequently asked questions)[61]
- Présentation du service
  entreprise support
  technologie utilisée
  étymologie du nom du service[62]
- Lien vers un (des) article(s) scientifique(s) publié(s) sur le système
  article descriptif (par les concepteurs de l'outil)
  article comparatif ou critique (par un auteur extérieur)
- Message d'erreur si le traitement de la requête échoue[63]
- Adresse de courrier électronique pour contacter l'administrateur

IV.B.3. Historisation

IV.B.3.a. Marquage des documents déjà vus[64]

IV.B.3.b. Répétabilité et renouvellement

- Déterminisme[65]
- Exploration de la base, côté ludique et attractif, en faisant jouer le hasard : « roulette »[66], tirage aléatoire de documents dans un domaine (large) donné[67]

[58] La plupart des conseils dans ce domaine vont dans le sens d'une recherche moins restrictive. Mais par exemple utiliser astucieusement une notation avec un joker, ou utiliser des opérateurs moins contraignants que d'autres, sert d'abord à factoriser toute une combinatoire de termes ou de requêtes. La vocation première de ces fonctionalités est la notation synthétique d'une série de variantes, ce n'est donc pas la dégradation progressive de la précision !
C'est pourtant la vision que voudrait donner OPEN TEXT INDEX (cf. **IV.C.1.c.**)

[59] « If your search is in English, search for both British and American spellings. For example, you could search for labour OR labor. » [OPEN TEXT INDEX]

[60] En général, il s'agit de ne pas confondre les guillemets simples et doubles (‘ et « »), et de ne pas dissocier les éventuels opérateurs unaires + et - des termes sur lesquels ils portent (pas de blanc).

[61] Le ton y est informel. On y trouve même quelques pointes d'humour, par exemple chez HOTBOT :
« My browser crashed, what did HOTBOT do ? - HOTBOT's generally a good fellow. It's unlikely he caused you any harm. HOTBOT follows several rules, such as the Standard for Robot Exclusion [cf. **V.A.2.b.**]. [...] Rumors of HOTBOT eating children are greatly exagerated. »
« HOTBOT didn't return queries in the order I wanted, why ? - Although we have tried to create a ranking formula that will display our results in a meaningful order, we can't always know what you are thinking. (Mind reading will be included in Version 3.0.) »

[62] Par exemple :
« In the spirit of exploring and charting new worlds, MAGELLAN takes its name from Ferdinand Magellan, a Portuguese explorer who navigated the Strait of Magellan in 1520. »
Lycos avait aussi une jolie explication, mais qui est devenue introuvable depuis la mutilation sauvage de son aide-en-ligne.

[63] Les browsers Internet (Netscape, Mosaic...) fournissent des messages d'erreurs lorsque qu'une URL ne peut être atteinte (c'est le cas d'une indisponibilité momentanée -surcharge- ou définitive -disparition du lien-).

[64] Le marquage des liens déjà activés par un utilisateur donné est réalisé d'office par les browsers Internet.

[65] La même requête, lancée sur la même base (dont le contenu est inchangé), redonne les mêmes résultats. Dans la pratique, il faut savoir cependant que le contenu des bases se renouvelle et grandit sans cesse.

[66] « Play WebRoulette : If you're looking for adventure, play WebRoulette and Spidey will pull 10 URLs at random from our extensive index. You'll never know what you'll discover. » [WEBCRAWLER]

[67] Les moteurs de recherche associés à des catalogues classant les documents disposent d'une division de leur corpus en domaines thématiques, qu'ils peuvent utiliser ici.

IV.B.3.c. Personnalisation
sauvegarde d'une combinaison des paramètres de recherche
pour une session (au fil des diverses requêtes)
pour une recherche ultérieure
sauvegarde d'une configuration
(en fonction du browser[68], de la puissance de la machine, etc.)

IV.B.3.d. Conservation d'une requête[69]
- Retour à la dernière requête pour modification / affinement
- Mise en commun des requêtes les plus populaires[70]

IV.B.3.e. Récupération du corpus résultat
- Sauvegarde par un marque-page ou « bookmark »
(pour un dépouillement différé par exemple)[71]
- Reprise comme nouvel espace de recherche[72]

IV.C. Critères d'évaluation sur le plan documentaire[73]

IV.C.1. Rappel

IV.C.1.a. Influence des caractéristiques du service
(On observe que pour une requête équivalente, les résultats des différents moteurs se chevauchent sans se recouvrir[74])
- Contenu et couverture de la base (définit l'espace de recherche)
- Codage des données et mode de traitement

[68] Les moteurs de recherche, pour être accessibles au plus grand nombre de personnes, sont conduits à prendre en compte la diversité des browsers utilisés comme interface WWW. Leur tactique est alors celle qu'explique ici HotBot :
« When you connect to the HotBot search engine, HotBot will identify your browser software and assess your browser's capabilities. HotBot dynamically adapts its output to the capabilities and idiosyncracies of your browser software, permitting the use of powerful features while remaining inclusive. » [HotBot]

[69] Nous n'avons pas trouvé de possibilité de mémorisation de stratégie. Cette fonctionalité est la base d'un service de Diffusion Sélective de l'Information : une requête sauvegardée fait office de profil et peut être relancée périodiquement sur la base. Le service de DSI à proprement parler se relance de lui-même périodiquement et ne propose que les documents nouveaux depuis la dernière interrogation.
Infoseek Guide, dans sa rubrique « Your News / personalize », propose un service approchant, non pas pour une requête par mots-clés, mais pour le choix individuel d'un petit nombre de rubriques dans une liste (recherche de type catalogue, et non index). A tout moment, l'utilisateur accède ainsi directement à la dernière mise-à-jour sur les sujets qu'il a présélectionnés, par exemple le cours de la Bourse pour telle et telle entreprise.

[70] Open Text Index entretient une sorte de palmarès des requêtes et des documents-résultats correspondants :
« Do you keep track of what people search for ? - We keep track of which words are searched for most often, and which pages show up as results most often. We thought this information was so cool that we decided to publish it on the Web so you can read it, too. You can read our WebPulse ! page, which is updated every business day. » [Open Text Index]
Derrière l'interrogation de l'utilisateur, on pourrait aussi lire une inquiétude quant au caractère privé et à la discrétion souhaitée pour ses requêtes personnelles. Ce genre de souci sera tempéré du fait de l'automatisation des systèmes de recherche et du volume de données manipulées (l'information personnelle est tout simplement noyée dans la masse). Il peut davantage être présent dans le cas de petites bases d'accès limité.

[71] Une recherche peut demander du temps, et la liste des documents sélectionnés est souvent longue : il est utile de prévoir le cas où l'utilisateur doit interrompre son travail de dépouillement des résultats.

[72] Les bases documentaires classiques (style INSPEC par exemple) permettent tout à fait la reprise de résultats de recherches antérieures comme sous-parties d'une nouvelle requête. Nous n'avons pas trouvé de telle fonctionnalité sur les outils de recherche sur WWW. AltaVista, dans ses requêtes, permet cependant qu'une première équation booléenne définisse un espace de recherche (« search field »), avant de sélectionner prioritairement les documents contenant certains mots donnés (les mêmes ou d'autres - « ranking field »).

[73] L'évaluation est d'autant plus délicate que l'utilisateur ne sait pas toujours très bien ce qu'il recherche, cf. [Paijmans]. Les critères formels sont alors inexploitables.

[74] La diversité des résultats d'un outil à l'autre amène souvent dans la pratique à systématiquement consulter plusieurs services pour disposer de résultats plus complets. Cette stratégie s'est traduite par l'apparition de « méta-services de recherche » : l'utilisateur soumet une requête, celle-ci est traduite et transmise à différents moteurs ; les résultats recueillis sont réorganisés en une seule liste pour l'utilisateur.
L'inconvénient à notre avis de ce système, tel qu'il est mis en œuvre actuellement, c'est qu'il interdit toute formulation un peu évoluée, car ne peuvent être utilisées que les options communes à tous les outils : on en reste à une requête standard et peu puissante. On pourrait concevoir un traitement plus nuancé, sans ce nivellement par le bas.

IV.C.1.b. Type de besoin de l'utilisateur
- Une partie des documents pertinents suffit, pour
  avoir une idée de la tendance générale, par un aperçu représentatif
  obtenir un renseignement donné : il suffit de trouver
  un document (fiable) le fournissant
  obtenir des points d'entrée sur WWW
  pour ensuite l'explorer grâce aux liens[75]
- Avoir tous les documents (vision exhaustive)
  recherche systématique et exhaustive : il importe de connaître chacun des documents

IV.C.1.c. Capacités cognitives
- Volume de résultats énorme,
  qui amène à préférer une vision sélective plutôt qu'exhaustive[76]

IV.C.2. Précision[77]

IV.C.2.a. Vision d'ensemble, synthétique
- Pôles, documents centraux (par exemple,
  si on soumet le nom d'une personne, sa page personnelle ;
  ou le nom d'une entité, sa page d'accueil (homepage))

IV.C.2.b. Vision pointue, analytique
- Documents les plus spécifiques à un sujet

IV.C.3. Rapidité de convergence

IV.C.3.a. Changements significatifs (sensibilité de l'indexation)

IV.C.3.b. Changements dans le bon sens (interprétabilité et prédictibilité des résultats)

V. Annexe : ce qui est lié au contexte applicatif

V.A. Politique de gestion de la base

V.A.1. Accessibilité

V.A.1.a. Large[78]

V.A.1.b. Libre ou contrôlée

[75] C'est exploiter le fait que l'on pointe sur une structure maillée, sous-jacente à l'espace de recherche et même le débordant.

[76] D'où l'argumentaire de OPEN TEXT INDEX :
« Can I do « fuzzy searching » with the OPEN TEXT INDEX ? - Fuzzy searching is a way of expanding the results of a search. This isn't usually necessary if you use the OPEN TEXT INDEX. You're more likely to find too much information than not enough. We focus, instead, on helping you refine your search ». [OPEN TEXT INDEX]
cf. aussi **IV.B.2.b.**, les conseils pour réduire le silence (et note).

[77] Vu la masse considérable de données d'une part, et l'habituelle pauvreté des requêtes d'autre part (elles se réduisent à quelques mots, pas toujours très bien choisis), la précision reste difficile à assurer ; elle pourra être très inégale, notamment en fonction de la couverture de la base dans le domaine du thème recherché.
[Koch] va jusqu'à dénoncer le peu de cas fait à la recherche d'une bonne précision :
« The services are in general set up for high recall. Instruments raising precision, like for instance relevance feedback (by EXCITE ; or not functioning correctly as by OPEN TEXT INDEX « find similar ») or vocabulary control, are seldom used. One service (EXCITE) offers query expansion, as default, in this case called concept-search. This option is however not affectable by the user and thereby set up for high recall. An equivalent by MAGELLAN (with PLS) has been removed. MAGELLAN, OPEN TEXT and EXCITE are offering, also directly, to focus the result with help of further search-arguments, although without sufficient support. » [Koch]

[78] Le choix de l'infrastructure de WWW, standard et utilisée universellement, ne nécessitant pas d'installation sur le poste client, facilite et généralise l'accès à ces services.

V.A.2. Protocoles de visite
V.A.2.a. Repérage des nouveaux
- Les nouveaux se déclarent eux-mêmes (par inscription) :
  sites visités sur demande
  fourniture d'une description selon une norme prescrite
- Sites recommandés[79]
- Suivi des liens des sites déjà connus et enregistrés[80]
  en profondeur d'abord
  (exploration récursive du premier lien hypertextuel rencontré)
  en largeur d'abord
  (exploration de tous les liens d'une page avant de descendre dans un des liens hypertextuels)
  en profondeur *p* et en largeur *l* limitées (par exemple *p*=3 et *l*=10)
- Consultation des listes des nouveautés

V.A.2.b. Savoir-vivre vs. intrusivité
- Respect d'interdictions d'accès
- Possibilité de se désinscrire
- Discrétion (pas de surcharge de la machine)[81]
- Indication qu'une visite a été faite

V.A.3. Montée en volume
V.A.3.a. Performance
- Repérage de nouveaux documents
- Traitement
- Stockage[82]
  (un codage volumineux doit se justifier par un gain suffisant en qualité)

V.A.3.b. Organisation
- Centralisée
- Distribuée, répartie[83]

[79] Ce peut être une forme de mise en commun de découvertes, suite à une recherche par un autre moteur, une navigation, etc.

[80] La stratégie de parcours des liens n'est pas neutre : un parcours en profondeur d'abord sera efficace pour repérer des documents détaillés et spécialisés ; un parcours en largeur d'abord favorise une bonne couverture générale, en recensant plus spécialement les pages d'accueil et en diversifiant les serveurs. Il reste que c'est, dans les deux cas, un pari sur la connexité du réseau WWW.

[81] Un parcours des liens en largeur d'abord, et qui privilégie la diversification des serveurs, peut faire des consultations moins longues auprès de chacun des serveurs.

[82] Le moteur de recherche utilisé par HARVEST BROKER, GLIMPSE, fait le choix de privilégier un petit encombrement mémoire, au prix d'un traitement légèrement plus long. Il écarte ainsi la solution de l'index, encombrant et rigide vis-à-vis des fautes de frappe.
« In some sense, GLIMPSE takes the opposite extreme to inverted files in the time vs. space tradeoff [...]. For some applications, such as management of personal information, speed is a secondary issue. Most users would rather wait for 10-15 seconds, or even longer, for a query than double their disk space. [...]
We call our method « two level searching ». The idea is a hybrid between full inverted indexes and sequential search with no indexing. It is based on the observation that with current computing performance, sequential search is fast enough for text of size up to several megabytes. [...] the index does not provide exact locations, but only pointers to an area where the answer may be found. Then, a flexible sequential search is used to find the exact answer and present it to the user. » [Manber]
Notons cependant que cette stratégie a d'abord été pensée dans le cadre de bases locales, personnelles, qui ne sont pas du même ordre de grandeur que WWW auquel elle s'attaque maintenant (via HARVEST BROKER).

[83] Les services de recherche ont une base de données centralisée, qui leur permet de situer chaque document dans le contexte d'ensemble ; mais certains prévoient à terme une architecture répartie, en raison de la difficulté de canaliser et traiter toute l'information en un seul point. Des avancées restent à faire, pour trouver le moyen de coordonner différentes bases de la manière la plus satisfaisante possible.

V.A.4. Intérêts commerciaux
V.A.4.a. Emergence (les documents qui apparaissent les premiers dans la liste des résultats ont une position privilégiée : l'attention du lecteur est plus vive, et d'autre part le lecteur ne persiste pas toujours à dépouiller plusieurs pages.)
- Contrecarrer les détournements
  pénalité pour répétition (« spamming »)[84]
- Contrôler et exploiter cette propriété
  vente de mots-clés[85]

V.B. Données disponibles : hétérogénéité des documents
V.B.1. Qualité de la forme[86]
V.B.1.a. Structuration
- Mauvaise utilisation du codage HTML
  document non conforme
  sous-utilisation de balises importantes pour la description du contenu
  confusion entre la structure fonctionnelle et la mise en page

V.B.1.b. Rédaction
- Titre représentatif vs. creux
- Vocabulaire
  pauvre
  imprécis

V.B.1.c. Contextualisation
- Rattachement
  possibilité de rapporter la page à l'ensemble auquel elle est initialement rattachée[87]
  caractère « auto-portant » (la page telle quelle, isolément, est compréhensible)
- Frontières
  possibilité (ou difficulté) de désigner un ensemble de pages comme définissant un document complet et cohérent
  différenciation des niveaux (l'indexation d'un serveur n'est pas du même ordre que celle d'un paragraphe)

V.B.2. Internationalisation
V.B.2.a. Langues[88]
V.B.2.b. Jeux de caractères
- Caractères diacritiques
- Alphabets différents (japonais, arabe, etc.)

---

[84] Les moteurs étant généralement sensibles aux fréquences des mots, certains ont voulu abuser de cette propriété en commençant leur document par des dizaine de fois le même mot-clé répété. Des illustrations spectaculaires de ce procédé peuvent être trouvées dans [Koch] -illustration 26 et suivantes ; le manuel de HotBot rapporte encore des ruses insoupçonnées :
« It has become popular for people to create pages that maliciously « spoof » search engines into returning pages that are irrelevant to the search at hand, or which rank higher that their relevance or content warrant. Common examples of spoofing are duplicating words thousands of times in comments or keywords, or icluding large number of « invisible » words in a tiny font, or in the same color as the background color of the page. »
S'étant aperçu de la chose, certains moteurs de recherche ont alors tenté d'éliminer ces documents peu scrupuleux en rajoutant un seuil, avec une règle du genre : si un mot-clé apparaît plus de sept fois dans un document, alors celui-ci est présumé coupable de manœuvres. Mais ceci n'est évidemment pas infaillible...
Ces règles peuvent intervenir plus ou moins tôt : pour l'admission du document dans la base ; au niveau de l'évaluation de la pertinence, lors du calcul de proximité avec la requête ; ou encore tout-à-fait en aval, pour l'ordre de présentation des résultats.

[85] Moyennant finance, le service assure à telle page de sortir dans les toutes premières pour des sujets la concernant.

[86] Tout ce paragraphe est largement inspiré par [Koch].

[87] Le contexte n'est pas toujours bien repéré : il arrive de manquer de liens rattachant le document à son contexte structurel quand on arrive par un lien externe ou le résultat d'une recherche par index : « The author often forgets that his/her publication, when read and indexed, is mostly torn out of its context, and that title and other metadata, and sometimes even large parts of the text ought to consider this fact. The original context will not always be possible to reconstruct from the fragments shown in the search service. » [Koch]

[88] Il n'y a évidemment pas de multilinguisme, au sens où l'on utiliserait un module de « traduction automatique ». Ce dont il s'agit, c'est de pouvoir calculer les documents proches d'une requête formulée en n'importe quelle langue : les documents retrouvés, du fait de l'utilisation de l'analogie entre les mots de la requête et ceux des documents, seront le plus souvent de la même langue.
Ce multilinguisme n'est peut-être pas aussi crucial qu'il n'y paraît, car la grande majorité des documents sur WWW est en anglais, langue de l'international.

V.B.3. Intérêt du contenu

V.B.3.a. Portée

- Intérêt personnel
- Intérêt local (par exemple document de travail interne à une équipe)
- Intérêt de courte durée

V.B.3.b. Lectorat, public visé

- Indicateurs pour les documents réservés à un public « adulte »[89]

[89] Comme dans la presse, il y a sur WWW un nombre significatif de documents de moralité ou de décence discutable. Certains s'empressent de décliner toute responsabilité :
« Note that some search results might contain adult-oriented material or material that you may find objectionable in nature. Please remember that these sites and newsgroups simply reflect what is currently available on the Internet. These documents are not created or maintained by us at INFOSEEK. » [INFOSEEK GUIDE]
D'autres ont créé un indicateur correspondant, mais qui suppose un comité de lecture, donc une analyse des documents non entièrement automatisée :
« MAGELLAN's Green Light appears next to reviewed sites that, at the time of review, contained no content intended for mature audiences. [...] Sites without the Light are not necessarily objectionable -they may simply contain adult topics or unregulated content. »

3. Points d'approfondissement

3.1. L'utilisation du document comme requête : QBE, relevance-feedback

Commençons par citer les passages relatifs à ce sujet, dans les aides en ligne de chaque outil successivement :

EXCITE : option « More like this »

« The « Find similar » feature : ICE [Intelligent Concept Extraction] is especially good at finding documents similar to a given document. So, if you [think] one of the documents in your result list is close to what you are looking for, click the « Find similar » link next to the title. For example, if you searched for « jordan », you would find documents about Air Jordan as well as the country Jordan. If you where actually interested in the country, click on « Find similar » next to one of the documents about Jordan country, and ICE will find more documents about the country, and none about the basketball star. »

« Find similar documents : So you've found the perfect page, and now you want more. Most other search engine would make you re-type your query, manually picking out terms to describe your document. With EXCITE's Query-by-Example, you can avoid the extra steps and look for similar pages with a single click. »

« Query By Example : When a user finds a document that contains relevant information, he or she may logically wish to use this information for additional searches. To automate this process, excite provides a means to 'click' on the document itself, and use its entire contents as search parameters to locate similar documents. »

INFOSEEK GUIDE

« Find pages similar to a listed page : If you see a page that you're interested in, you can search for similar pages by clicking « Similar pages ». INFOSEEK GUIDE uses information about the [selected] page to search for [other] pages with similar content. Clicking on « Similar pages » will search all web pages (categorized and uncategorized). This takes you outside of the collection of INFOSEEK Select pages. »

OPEN TEXT INDEX : option « Find similar pages »

« What types of searches can I do with the OPEN TEXT INDEX ? - You can [...] search for pages similar to the first ».

« L'algorithme est simple et repose sur le nombre d'occurrences de mots présents dans le document. » [Plourde]

WEBCRAWLER : option « Find similar pages »

« The detailed format [for viewing results] provides titles plus summaries, URLs, numerical relevancy scores, and the option of viewing similar pages for each result returned. » [WEBCRAWLER]

WEBCRAWLER : option « Search the Web for more like... »

« When you're browsing through the WEBCRAWLER Select reviews, and find a subject or review that piques your interest, you can click on the Spidey Search icon to search the Web for related pages. Spidey search is brought to you by WEBCRAWLER editors who create special similarity queries for each Select subject and review, carefully choosing keywords that will produce optimal results. »

Ce dernier extrait expose un procédé fondamentalement différent des précédents, à savoir que la caractérisation du texte n'y est plus automatique. Ceci sort de notre champ

d'investigation : nous ne nous occuperons pas de cette option-là de WebCrawler, ni d'autres du même type.

QBE est une notation pour « Query-By-Example » : la requête n'est alors pas la donnée de quelques mots-clés, mais un texte représentatif des documents que l'on souhaite trouver. Le traitement associé est une comparaison de texte à texte, pour identifier les textes analogues au texte-requête.

Ce qui est mis en valeur dans la présentation faite de cette fonctionnalité par les services de recherche, c'est la simplicité d'utilisation. Un simple clic sur un document pertinent permet de relancer la recherche.

Cependant, tout n'est pas aussi simple qu'il n'y paraît, pour la suite des opérations du moins. Une requête-texte n'est pas de même nature qu'une requête par quelques mots-clés. Elle fournit a priori plus de termes de recherche (d'où souvent l'impact de la longueur du document). De plus, ces termes se comportent différemment de mots-clés choisis : la plupart, pris individuellement, sont moins précis, moins centrés sur le sujet, moins directement informatifs ; mais lorsque les termes sont considérés ensemble, ils se désambiguïsent réciproquement et décrivent plus complètement le sujet et son contexte.

On comprend donc que l'utilisation de QBE est délicate : pour être intéressante, il convient de prévoir un mécanisme adapté, spécifique, pour la construction de la requête à partir du texte.

Peu d'études ont été menées sur les requêtes-texte, aussi les systèmes existants sont-ils tentés d'appliquer au texte les mêmes procédures qu'aux mots-clés. Dans de tels cas les résultats obtenus ont été si peu satisfaisants que la fonctionnalité QBE a été purement et simplement supprimée (c'est sans doute ce qui est arrivé pour OPEN TEXT). De même, la qualité des résultats de QBE pour EXCITE et INFOSEEK GUIDE reste controversée : ces services veulent-ils à tout prix garder cette fonctionnalité rare et distinctive, au risque de décevoir les utilisateurs ? ou se mettent-ils à développer des modes de caractérisation des textes pour améliorer leur offre actuelle ?

Reste enfin à préciser la distinction entre QBE et relevance-feedback :

- le QBE permet de prendre un texte en guise de requête. Mais pourquoi en rester aux textes des documents fournis comme résultats d'une requête antérieure ? Il peut être intéressant de soumettre le texte que l'on veut : en faisant un copier/coller dans une fenêtre de requête prévue à cet effet, en donnant une adresse http, etc.
- le relevance-feedback désigne les procédés permettant de prendre en compte l'évaluation des résultats par l'utilisateur. Suite à une proposition de documents en réponse à une requête, il s'agit d'affiner la recherche, en retrouvant d'autres documents du même style que ceux jugés effectivement pertinents, et/ou en écartant des propositions analogues à celles des documents rejetés. Mais garde-t-on bien, comme base de la nouvelle requête, la requête initiale, qui indique le thème central de la recherche ? Il semble au contraire que dans les systèmes étudiés, la relance de la recherche par l'intermédiaire de documents se comporte comme une recherche indépendante, ne s'inscrivant pas à part entière dans l'élaboration d'une stratégie de recherche.

L'amalgame entre QBE et relevance-feedback a, on le voit, appauvri ces deux fonctionnalités :

- le QBE doit pouvoir considérer un document non nécessairement issu d'une requête antérieure ;
- le relevance-feedback se conçoit comme inscrit dans une stratégie de recherche, stratégie se déployant à partir de la requête initiale (et non pas « amnésique »).

3.2. La recherche par concepts

EXCITE est actuellement (et il s'en flatte) le seul moteur à proposer une recherche « par concepts ». Voici ce qu'il en est dit :

EXCITE

« EXCITE's search technology : Central to the technology used by EXCITE is the method of Intelligent Concept Extraction (TM), or ICE searching. This technological breakthrough is the result of significant revisions to EXCITE's highly successful, proprietary concept-based technology. Using our own advanced statistical algorithms, we are able to find and score documents based on a correlation of their concepts, as well as actual keywords. In effect, this method equals the dramatic improvements in both Recall and Precision achieved by Latent Semantic Indexing, but uses an entirely different technology which is far less demanding in computational power and therefore both rapid and scalable. Most users who have compared a variety of services agree that Excite is the clear leader in correctly discovering and ranking relevant documents, even if they fail to contain original keywords.

While the detailed operation of our patent-pending technology is confidential, EXCITE uses proprietary techniques that allow us to achieve the same sort of analysis as acomplished by Latent Semantic Indexing, but much more efficiently. »

« There are various methods of building concept-based indices, some of which are highly complex, relying on sophisticated linguistic and artificial intelligence theory that we won't even attempt to go into here. EXCITE sticks to a numerical approach. EXCITE's software determines meaning by calculating the frequency with which certain important words appear. When several words or phrases that are tagged to signal a particular concept appear close to each other in a text, the search engine concludes, by a statistical analysis, that the piece is « about » a certain subject. » [Barlow]

La recherche conceptuelle est séduisante de prime abord. Elle correspond à cette idée, que l'analyse linguistique doit jouer sur deux plans, pour retrouver, à partir de la forme de *surface*, la représentation *profonde* de ce qui est exprimé[90]. Les concepts exprimeraient le contenu même de l'information communiquée, alors que la recherche sur quelques mots susceptibles d'apparaître dans le texte touche très vite à ses limites avec les phénomènes de synonymie[91], de polysémie[92], d'implicite[93].

Qu'en est-il, concrètement, de ces concepts si prometteurs ? A savoir, comment sont-ils mis en œuvre, et, à travers cette modélisation, que représentent-ils, que signifient-ils « par construction » ?

Les concepts d'EXCITE semblent être tout simplement des *classes* de mots apparentés à un même thème.

Ces classes sont constituées automatiquement en fonction de la distribution des mots sur un corpus de documents : des critères statistiques permettent d'associer les mots entre eux. La

[90] Cf. aussi la théorie, généralement admise, qui articule et oppose le mot, la chose, et le concept. Elle est contestable dans la mesure où elle prive la langue de sa dimension sémantique, en plaçant la compréhension et l'interprétation dans un jeu autonome sur les concepts. La langue est reléguée au rang d'outil, simple véhicule pour « transporter » les idées, décodées puis traitées au niveau conceptuel. Des linguistes, comme François Rastier, nous mettent en garde contre ce réductionisme.

[91] Synonymie : je fais par exemple une recherche sur le sujet de la « diffusion ciblée » d'informations, et l'auteur d'un document qui m'aurait intéressé aura choisi d'employer un autre terme pour parler du sujet, comme « diffusion sélective ».

[92] Polysémie : ma requête porte sur la « diffusion »d'informations, et le système me retrouve des documents parlant des phénomènes physiques de « diffusion » de la chaleur, des gaz ou du rayonnement par exemple.

[93] Implicite : dans le cadre d'une recherche sur la « diffusion ciblée d'informations », je recherche les destinataires les plus concernés par un document ; il est donc évident que cela recouvre une étude de la pertinence, même si je ne l'explicite pas.

seule finesse peut-être prévue, c'est la possibilité pour un mot d'appartenir à plusieurs classes, traduisant ainsi des phénomènes de polysémie ou d'homonymie.

On peut imaginer que, une fois ces classes construites, le traitement d'une requête est assez sommaire : chaque terme renvoie à la ou les thématique(s) associée(s), ce qui permet de passer d'une requête par mots à une requête par thèmes ou « concepts »[94]. On néglige ou évince les alternatives thématiques qui s'écartent de la thématique dominante (opérant de la sorte une certaine « désambiguïsation »). Les mots de cette thématique dominante (ceux qui apparaissent dans le texte mais aussi les mots voisins, appartenant à la même classe élue) prennent alors un rôle prépondérant dans le calcul de la pertinence.

Le passage par les classes thématiques explique donc la possibilité de retrouver des documents dans lesquels aucun terme de la requête initiale n'apparaît, de même que la désambiguïsation contextuelle opérée.

Computationnellement parlant, le procédé est aussi avantageux, puisqu'il permet de passer d'un espace des termes à *n* dimensions (*n* = nombre de termes d'index) à un espace des concepts à *m* dimensions (*m* = nombre de classes), avec $m << n$.

L'idée, si elle n'est pas nouvelle, est astucieuse. Il est dommage d'en masquer la simplicité derrière l'appellation pompeuse de « recherche conceptuelle ». D'autant que les problèmes les plus intéressants résident dans la manière de construire et d'utiliser les classes thématiques, et que cela n'est peut-être pas si au point que cela dans les systèmes actuels.

3.3. Les modes de pondération : pondérations quantitative et qualitative

EXCITE se démarque encore des autres moteurs de recherche, en proposant depuis peu l'introduction de pondérations quantitatives :

EXCITE

« Avanced EXCITE search features : [...] Increasing a word's weight with ^. By adding a ^ symbol and a value to the end of a word, you can increase it's weight relative to the other words in the search and move documents containing it higher in the results list. Example search : dog care grooming^3 »

« Using the ^ symbol to increase a word's relative weight. *What it does* : EXCITE will view the word as more important to you in deciding what results to return to you. *How to do it* : Just follow the given words with a ^ symbol and the relative weight value. This can be any number as the values will be relative to each other. No value of course assumes a value of '1'. For example, college^3 footbal scores^6 will look for pages that contain those words, with an emphasis on 'college' and an even heavier emphasis on 'scores'. *What's different about it* : EXCITE uses the weights to determine what info is most important to you in returning your results. Normally EXCITE assumes all the words are equally important. *The downside* : None really, except that using weights on a lot of words can get cumbersome and you may supress good results if you are not careful. »

La première chose à remarquer, c'est que, sauf explications détaillées, l'utilisation de pondérations numériques est ambiguë, car il y a plusieurs manières de l'interpréter. Donnons différentes propositions toutes vraisemblables :

[94] Le passage des mots-occurrences aux concepts est un double mouvement de réduction et d'expansion. La réduction est celle qui fait que *n* mots différents se retrouvent étiquetés par un même concept ; la donnée du concept ne permet pas de savoir quels étaient ses représentants effectivement présents dans le texte ; la réduction opère une perte d'informations, elle rend donc la procédure irréversible. Mais le passage aux concepts effectue en même temps une expansion : le concept permet de prendre en compte non seulement les termes utilisés par l'auteur dans le texte, mais aussi tous les mots voisins, reformulations variées pour la même thématique.

- première proposition : les poids ont valeur de cardinaux relatifs. Ainsi (en adoptant les notations de EXCITE), « a^2 b^4 c^10 » est équivalent à « a b^2 c^5 », seul importent les rapports des poids deux à deux.
- deuxième proposition : les poids ont valeur de cardinaux absolus. En reprenant l'exemple précédent, on n'a plus l'équivalence entre « a^2 b^4 c^10 » et « a b^2 c^5 », parce qu'on aura associé à certaines valeurs de pondération un traitement particulier. Par exemple, à la pondération 1 correspond un rôle limité (« ce terme-là n'est pas mis en valeur »[95]), aux pondérations 2, 4, 5, un rôle équivalent et moyen, à la pondération 10 une influence décisive (par exemple la présence obligatoire du terme).
- troisième proposition : les poids ont valeur d'ordinaux. Pour la sélection de documents potentiellement pertinents, tout se passe comme s'il n'y avait pas de pondérations. Celles-ci n'interviennent qu'au moment de la présentation des résultats, pour les ordonner, en faisant passer en premier les documents qui comportent les termes les plus fortement pondérés. Ainsi, « a^2 b^4 c^10 » et « a b^2 c^5 » seraient équivalents ; on sélectionne les documents ou apparaissent a, b, ou c ; on présente d'abord ceux qui présentent a, b et c, puis b et c, puis a et c, puis a et b, puis c, puis b, puis a[96].

On peut imaginer bien d'autres propositions, par combinaison des précédentes ou encore différentes.

Les pondérations quantitatives ont l'avantage d'être d'utilisation (illusoirement) simple, mais leur sémantique effective est souvent laissée dans le vague (cf. les propositions ci-dessus), leur affectation est arbitraire (pourquoi choisir une pondération de 4 plutôt que de 5, par exemple ?), et de toutes façons elles ne sont qu'un moyen de calcul, étranger à la nature linguistique et textuelle des données.

Ces critiques nous invitent à rechercher des formes de pondération plus qualitatives.

Une fonctionnalité maintenant assez courante des systèmes de recherche est la mise en valeur d'un terme, en indiquant que sa présence est exigée dans les documents résultats.

C'est souvent le signe +, accolé aux termes requis obligatoirement, qui représente cette option.

Mais c'est hélas à peu près tout. On aurait pu imaginer d'autres formes de mise en valeur moins brutales. Par exemple, cela aurait pu être le cas de l'attribut « should », dans HOTBOT :

> « Choosing « should » tells the search engine to place more emphasis on documents which contain the specified word than those which do not. The engine will still return documents which don't contain the word or phrase, but will rank them lower on the list than those that do. » [HOTBOT]

Mais ici « should » est la seule option possible entre les deux extrêmes « must » et « must not » ; il ne s'agit donc pas vraiment d'une préférence marquée par rapport à un autre terme de la requête qui lui serait « neutre ». On ne fait que retrouver l'heuristique selon laquelle plus un document comporte de termes de la requête, plus il est pertinent.

Une autre manière de conférer une certaine importance à un terme est de lui donner un rôle renforcé dans le classement des résultats retournés : cf. **IV.A.4.c.**

[95] De même, si on a affaire à un texte et que les termes sont pondérés par leur nombre d'occurrences, les termes d'occurrence 1 (dits « hapax ») peuvent être considérés comme accidentels par rapport au sujet du texte. D'ailleurs, le fait d'avoir un terme (« hapax ») pour les désigner spécialement, montre bien qu'on leur accorde un statut bien particulier.

[96] Il n'y a pas de manière unique et évidente de définir un ordre sur les classes de résultats, car on a en fait deux ordres totaux à concilier et à équilibrer : celui du nombre de termes présents, et celui de la valeur des poids. Par exemple, quand on aurait une requête de la forme « a b c e^3 », donnera-t-on d'abord un document qui contient a, b et c, ou un document qui ne contient que e ?

Ces questions de modulation de l'importance accordée à un terme autrement que par le biais de pondération chiffrée (arbitraire, relative mais aveugle , monodimensionnelle) serait une piste à creuser davantage : cela développerait la notion de pondération qualitative, par opposition à un système purement quantitatif. On voit cependant que cette piste n'est qu'à peine explorée par les outils actuels.

4. Conclusion

L'utilisation courante de moteurs de recherche sur WWW révèle des disparités d'un outil à l'autre. De fait, chaque étape du traitement, depuis la couverture de la base de références jusqu'à la modélisation de la pertinence, est un passage déterminant pour le contenu et l'interprétabilité de la réponse. Et l'analyse détaillée que nous avons menée sur les différents paramètres de modélisation montre que, si la plupart de ces paramètres sont connus depuis longtemps, il n'y a pas (pas encore ?) de proposition consensuelle et optimale pour leur mise en œuvre conjointe.

Une démarche méthodique consisterait maintenant à définir la sémantique de tous les paramètres recensés, à repérer leurs interrelations, afin d'en proposer une combinaison moins empirique et plus maîtrisée. Cela s'inscrirait dans un cadre général, ouvert à (et peut-être suggérant) l'introduction d'autres paramètres. Autant les systèmes de recherche par plan de classement sont figés, autant, dans un système automatique à base d'index, des améliorations sont toujours envisageables.

La présente étude va ainsi être mise à profit à la DER pour le développement de la nouvelle version de l'outil DECID : en effet, sa fonction d'aide à la Diffusion Ciblée fait appel à une comparaison automatique de textes caractérisés par une indexation.

5. Bibliographie-Webliographie thématique commentée

5.1. Comparatifs : évaluation des systèmes

BARLOW Linda (1996) - « The spider's apprentice : How to use Web search engines »
http://www.monash.com/spidap
Comme son titre l'indique, il s'agit d'une introduction aux moteurs de recherche, un guide pour leur choix et leur utilisation. Astucieux et bien fait, ce document ne sera pas utile qu'aux débutants ! Il est apparemment bien mis à jour (bon suivi de l'actualité) : une rubrique résume les principales nouveautés et les tendances actuelles. Tout ceci sur un ton informel, facile et agréable à lire.

CONTE Ron (1996) - « Guiding Lights », *Internet World*, May 1996, pp.40-44.
Pour un rapide mais efficace panorama des principaux outils de recherche sur WWW : moteurs de recherche, catalogues thématiques, signalement des nouveautés, répertoires d'adresses e-mail, archives Gopher, logiciels en libre disposition sur le réseau, News Groups, interfaces regroupant et unifiant les recherches de différents moteurs.

PLOURDE Jean-Noël (1996) - « Critère et évaluation d'outils de recherche des ressources dans Internet », *Cursus* [périodique électronique étudiant de l'EBSI de l'Université de Montréal], Vol.1 n°2.
http://mistral.ere.umontreal.ca/~beaudryg/cursus/vollno2/plourde.html
Un article sérieux, d'une excellente revue. L'organisation générale de l'étude est intéressante : le plan offre à notre avis un parcours rationnel de la problématique. Au niveau des détails, beaucoup d'informations sont rassemblées, mais elles ne sont pas toujours aussi claires et précises que nous l'aurions souhaité : cela reste quelquefois de simples indications, difficiles à traduire de façon exacte et rigoureuse pour notre grille systématique.

SULLIVAN Danny (1996) - « The Webmaster's guide to search engines and directories »
Voir en particulier les liens :
- Features chart
- How search engines work
http://calafia.com/webmasters/
Faire recenser et indexer son site ou sa page WWW donne des résultats d'autant plus satisfaisants que l'on comprend le fonctionnement des différents moteurs de recherche et en particulier le traitement appliqué aux documents. C'est la perspective prise ici par l'auteur, celle des « webmasters who care about how search engines catalog their sites ». « How search engines work [?] [...] In fact, usually search engines produce so many relevant results that it is difficult to understand why a page ranked first did better than another page ranked 20th. This is the key question for the webmaster : why are some pages making it to the top of the list while others aren't. » Ce document a plusieurs points forts : clarté, bon sens, beaucoup d'éléments d'information de sources variées, bonne mise à jour. L'approche adoptée est plus empirique et moins systématique que la nôtre.

VENDITTO Gus (1996) - « Search Engine Showdown : IW Labs tests seven Internet search tools », *Internet World*, May 1996, pp.78-86.
Un article assez complet et plein de bon sens, souvent donné en référence. Dans un style rédigé et construit, il souligne les caractéristiques les plus saillantes de chaque moteur, du point de vue de l'utilisateur : c'est agréable à lire et instructif, sans avoir une approche très systématique. L'argumentaire est basé sur l'analyse des résultats de trois requêtes : « NBC » (très grande entreprise, donc apparaît dans beaucoup de documents mais pas toujours avec la même pertinence) ; « Steve Forbes » (homme politique américain actuel, donc requête sensible aux mises à jour, par exemple pour trouver son dernier emploi du temps) ; « Internet access via satellite » (requête où plusieurs concepts sont en relation).

WINSHIP Ian R. (1995) - « World Wide Web searching tools - an evaluation »
http://www.bubl.bath.ac.uk/BUBL/IWinship.html
Les moteurs de recherche sur WWW sont ici étudiés au regard des fonctionnalités offertes par les grandes bases documentaires sur CD-ROM. Bien que ce document commence à dater (juin 95), une bonne part des remarques restent tout-à-fait pertinentes. Ainsi, il pointe avec justesse les principales défaillances des interfaces proposées. D'autre part, les tableaux comparatifs fournissent une bonne base de critères, équilibrée et organisée. Enfin le choix des requêtes d'évaluation est astucieux (une requête très pointue ; une autre dans un domaine non académique ; une troisième pour un sujet non américain). L'interprétation des résultats correspondants est faite avec finesse, évitant des conclusions hâtives pourtant tentantes.

5.2. Cadre d'analyse et critères généraux d'évaluation

KOCH Traugott (1996) - « Internet search services »
`http://www.ub2.lu.se/tk/demos/DO9603-meng.html`
Un document de référence ! Fait remarquablement le point des forces et des faiblesses des types de services de recherche offerts sur WWW, en prenant bien en compte les spécificités de ce réseau mondial. On en tire des enseignements sur les orientations à prendre ou à éviter, et sur les évolutions qui se dessinent. Ainsi, les catalogues, du fait de leur construction manuelle, ont à la fois une (forte) valeur ajoutée (sélection, structuration) et une couverture limitée (volume, mise à jour). S'ils sont utiles pour donner des points de départ de navigation, ils gagneraient à être plus étroitement combinés aux outils de recherche par index.
Cette analyse est réalisée dans le cadre du projet européen DESIRE, dont la présentation fait l'objet d'un chapitre. Rédigé initialement en langue allemande, le texte, transposé en anglais, comporte quelques expressions, ponctuations et tournures maladroites ne facilitant pas la lecture. Mais souhaitons qu'elles ne découragent pas le lecteur ! En revanche, il comporte toute une collection d'exemples illustratifs très parlants (sous forme de copies d'écran), et abondance d'autres liens utiles (avec un volume considérable de références, difficile à tenir à jour).

KOCH Traugott (1996) - « Search services : analytical form »
`http://www.ub2.lu.se/desire/radar/search-services/analyt-form.html`
Une des grilles d'analyse les plus complètes et les plus équilibrées. A partir d'elle, a été réalisée toute une série de fiches décrivant une vingtaine de moteurs de recherche, que l'on trouve à partir de l'adresse suivante :
`http://www.ub2.lu.se/tk/websearch_systemat.html`

PAIJMANS Hans (1992) - « An inventory of models in information retrieval »
`http://pi0959.kub.nl:2080/Paai/Onderw/Paai/Ai_ir/ai_ir.html`
Un point de vue intéressant, critique et constructif, pour un lecteur déjà familiarisé avec le domaine. Affine et nuance les concepts les plus utilisés : le document et sa représentation ; le rôle de la réduction dans la description ; les besoins -parfois imprécis- de l'utilisateur (typologie des requêtes, exigence ou non d'exhaustivité, tactiques, contrôle de l'interprétation faite par le système de sa requête) ; la prise en compte de l'auteur ; l'écart entre « document » et « information ».

5.3. Sur les techniques de calcul implémentées

EXCITE Inc. (1996) - « Information Retrieval technology and Intelligent Concept Extraction (TM) searching »
`http://www.excite.com/ice/tech.html`
Morceau de bravoure en tant qu'exposé développé, brillant et soigneusement rédigé (tout en étant dans le cadre d'une aide en ligne), il ne s'agit pas d'un article scientifique, mais d'un document dont le ton général est beaucoup moins sympathique que celui de [Pinkerton]. L'« Intelligent Concept Extraction » (ICE) utilisé dans EXCITE, seul service à revendiquer une telle « indexation conceptuelle », est lourdement vanté, qui plus est avec des formulations creuses, imprécises ou nimbées de mystère. C'est décevant ! Les principes sous-jacents sont tout juste décodables par les lecteurs experts dans le domaine. Cette volonté manifeste de ne pas communiquer d'information sur la stratégie appliquée apparaît encore dans l'accès à cette page, très indirect et difficile à trouver depuis la page de lancement des requêtes EXCITE. Reste une présentation synthétique des modèles de référence employés pour la recherche documentaire automatisée : l'exposé a alors des qualités pédagogique évidentes, même s'il reste un peu grossièrement orienté. D'où finalement trois raisons de le retenir dans notre Webliographie nonobstant nos reproches : (i) son état de l'art élégant ; (ii) son positionnement stratégique, comme document le plus informatif (!) sur l'indexation conceptuelle pratiquée par EXCITE (pour l'instant unique en son genre sur WWW) ; (iii) son exemplaire dissimulation d'information.

PINKERTON Brian (1994) - « Finding what people want : Experiences with the WEBCRAWLER »
`http://info.webcrawler.com/bp/WWW94.html`
C'est un article d'une dizaine de pages, agréable à lire, suffisamment précis et clair dans ses explications (chose rare, donc d'autant plus appréciable), et présentant plusieurs idées intéressantes. Ces qualités justifient qu'il soit signalé ici, même si l'outil associé (WEBCRAWLER) n'est pas dans les tout premiers en termes de popularité. Dans la première moitié de son article, l'auteur expose le mode de fonctionnement de WEBCRAWLER ; les principes utilisés à la base des algorithmes sont explicités. La deuxième moitié rassemble les « observations et enseignements » tirés de 6 mois d'exploitation du service, et notamment d'échanges avec les utilisateurs (courrier spontané, enquête avec questionnaire).

MANBER Udi, WU Sun (1993) - « GLIMPSE : a tool to search through entire file systems »
Le fichier postscript est accessible depuis
`http://harvest.transarc.com/afs/transarc.com/public/trg/Harvest/papers.html`
GLIMPSE est le moteur de recherche de HARVEST BROKER et fonctionne typiquement sur la base d'expressions régulières. L'article s'intéresse ici au stockage et à l'accès des données. Les fichiers inverses permettent des

recherches rapides, mais sont volumineux et ne sont pas adaptés à des expressions régulières. La proposition des auteurs est d'enchaîner d'abord une recherche par un petit index, qui ne fait que localiser les zones (larges) d'occurrence des termes ; puis, dans ces zones, c'est une recherche séquentielle qui est faite (telle celle d'une commande Unix « grep »). Ce qui est intéressant dans le cadre de notre étude, c'est l'analyse des procédés de recherche en deux classes (séquentiel vs. indexé), et la question des recherches sur des chaînes avec variantes (fautes d'orthographes, expressions régulières).